International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

EYESWIDE: Proctoring System for Online Examinations

1st Om Patil 2nd Aditya Shinde 3rd Aditya Tripathi 4th Shubham Wagh
Dept. of Artificial Intelligence and Data Science
Datta Meghe College of Engineering
Navi Mumbai, India
ompatil6948@gmail.com iamad1tyashinde2004@gmail.com adityastripathi28@gmail.com waghshubham197@gmail.com

5th Ms. **Poonam Kamble** (Project Guide)
Dept. of Artificial Intelligence and Data Science
Datta Meghe College of Engineering
Navi Mumbai, India
poonam.kamble11992@gmail.com

***Abstract*—The rapid growth of online education has introduced significant challenges in maintaining academic integrity during remote examinations. This paper presents "EYESWIDE," an AI-based online exam proctoring system designed to detect suspicious activities in real time. The system integrates computer vision and audio analysis techniques, including face detection, gaze tracking, head pose estimation, and object detection. A temporal analysis approach using LSTM networks is employed to evaluate behavioral patterns over time, reducing false positives and improving detection accuracy. Experimental results demonstrate that the proposed system achieves high accuracy with low false alarm rates. The system is efficient, scalable, and privacy-aware, making it suitable for modern online examination environments.**

***Index Terms*—Artificial Intelligence, Online Exam Proctoring, Computer Vision, Deep Learning, Gaze Tracking, Fraud Detection**

I. Introduction

The proliferation of digital learning platforms, asynchronous degree programs, and massive open online courses (MOOCs) has fundamentally restructured the contemporary pedagogical landscape, democratizing access to high-quality education across vast geographic and socioeconomic boundaries [1]. However, this monumental paradigm shift introduces a significant operational and ethical bottleneck: the secure, scalable, and fair administration of high-stakes examinations [3]. The foundational integrity of academic credentials inherently relies upon the verifiable assurance that the submitted assessment accurately reflects the unassisted cognitive competence of the authenticated student. In entirely unproctored online environments, instances of academic dishonesty—ranging from localized peer collaboration and the surreptitious use of unauthorized electronic devices to the deployment of highly sophisticated generative AI models and browser extensions—have demonstrably proliferated [5]. Recent empirical studies suggest that detected online cheating cases have surged by over fifty percent since the widespread adoption of remote learning protocols in early 2020, highlighting an urgent, industry-wide imperative for intelligent, automated proctoring solutions [5]. Historically, educational institutions and certification bodies have relied upon conventional onsite testing centers or synchronous live remote proctoring. Live remote proctoring models necessitate a human invigilator actively monitoring continuous video and audio feeds in real time [7]. While demonstrably effective at deterring flagrant misconduct, this approach suffers from severe and inherent scalability constraints, prohibitive financial costs per examinee, and the psychological and cognitive fatigue associated with human reviewers attempting to monitor multiple simultaneous feeds [9]. Consequently, the educational technology sector has witnessed the rapid conceptualization and deployment of fully automated, Artificial Intelligence-based Online Exam Proctoring (OEP) systems designed to autonomously flag anomalous behavior [8].

The proposed system, formally designated as EYESWIDE, addresses the critical, unresolved gaps present in contemporary, commercially available OEP solutions. A significant majority of existing commercial systems rely heavily upon static, heuristic-based anomaly detection mechanisms, which frequently generate an overwhelming volume of false positives—such as indiscriminately flagging a student for looking away from the screen to briefly ponder a complex mathematical problem or to utilize authorized scratch paper [12]. These algorithmic false positives not only place an undue administrative burden on human reviewers required to audit the flags but also significantly elevate the baseline anxiety of the test-taker, inadvertently depressing authentic academic performance and introducing construct-irrelevant variance into the assessment [4]. Furthermore, existing architectural models predominantly operate under the flawed assumption of ubiquitous high-bandwidth availability, requiring the continuous streaming of raw, uncompressed video data to centralized cloud servers. This approach poses severe cybersecurity and privacy risks

while actively marginalizing students residing in resource-constrained or rural regions characterized by high-latency internet connections [4].

EYESWIDE systematically mitigates these technical and ethical challenges through the deployment of a hybrid, multi-layered computational architecture. The framework synergistically combines a robust, client-side browser lockdown protocol with an edge-optimized, multi-modal AI processing engine. The system continuously evaluates an array of visual cues, encompassing biometric identity verification, precise gaze estimation, dynamic head pose calculation, and prohibited object localization, alongside acoustic signals designed to detect unauthorized speech or anomalous environmental noise [16]. Crucially, rather than relying on isolated, instantaneous frame analysis, the EYESWIDE architecture employs sequence-aware machine learning models—specifically Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. These temporal models rigorously analyze the sequential flow of human behavior, thereby calculating a highly dynamic "cheating probability score" predicated upon prolonged, contextual deviations from individualized normative baselines [18].

This comprehensive report meticulously details the architectural framework underpinning the EYESWIDE system, the mathematical and algorithmic methodologies of its constituent computer vision and audio processing modules, the specific datasets required for robust model training, and the indispensable ethical and privacy-preserving frameworks necessary for responsible deployment in modern educational institutions.

A. Contributions

The main contributions of this paper are as follows:

- Development of a multi-modal AI-based proctoring system combining computer vision and audio analysis.
- Integration of gaze tracking, head pose estimation, and object detection for enhanced monitoring.
- Use of LSTM-based temporal analysis to reduce false positives and improve detection accuracy.
- Implementation of edge-based processing to ensure privacy, efficiency, and scalability.

II. The Evolution of Remote Proctoring Systems

The evolution of remote assessment supervision can be systematically categorized into three distinct generational phases, transitioning from basic digital application restrictions to holistic, intelligent, and temporally aware behavioral analysis. An exhaustive review of the literature reveals the compounding nature of these technologies, wherein each generation attempts to resolve the specific vulnerabilities exposed by its predecessor.

A. First-Generation Systems: Application and Interface Restrictions

Early attempts to secure the integrity of online examinations relied almost exclusively on software-based environmental restrictions, commonly referred to within the industry as "Lockdown Browsers." These applications operate by interfacing directly with the host machine's operating system, modifying registry keys or actively terminating unauthorized background processes to construct a highly restricted "kiosk" mode environment [20]. Upon execution, the lockdown browser restricts the test-taker from navigating to unauthorized URLs, utilizing standard operating system keyboard shortcuts such as task switching or copy-pasting, accessing secondary monitors, or launching covert communication software including virtual machine hypervisors or remote desktop applications [20].

While highly effective at preventing on-device, purely digital cheating vectors, lockdown browsers are entirely blind to the physical environment encompassing the student. Test-takers can effortlessly bypass these digital constraints by consulting a secondary physical device such as a smartphone or hidden tablet, referencing concealed handwritten notes, or engaging in unauthorized collaboration with a peer located just outside the peripheral vision of the assessment hardware [22]. Furthermore, research indicates that the installation and execution of these intrusive root-level applications often induce severe technical complications, including system crashes on unstable network connections and fundamental incompatibilities with specific accessibility hardware or specialized operating systems like ChromeOS, leading some educators to abandon them entirely [20]. Thus, contemporary literature overwhelmingly concludes that software restrictions, while serving as a necessary foundational baseline, must be inextricably paired with advanced environmental monitoring to ensure comprehensive assessment security [24].

B. Second-Generation Systems: Heuristic-Based Environmental Auditing

To directly address the physical blindness of first-generation software restrictions, second-generation systems introduced integrated video and audio monitoring apparatuses. Initially, this paradigm involved continuous video recording architectures, often termed "Record and Review" models, wherein vast quantities of localized footage were stored and subsequently transmitted for asynchronous human auditing [26]. As the sheer volume of recorded assessments rapidly outpaced human reviewing capacity, these systems evolved into automated, heuristic-based frameworks that applied rudimentary computer vision algorithms to trigger administrative alerts based on rigid, hard-coded rules [28].

For example, early heuristic models dictated that if a Haar Cascade face detection algorithm failed to locate a human face within the frame for a duration exceeding three seconds, or if ambient audio levels exceeded a predefined, static decibel threshold, a violation flag was automatically generated and logged [28]. However, these rigid systems fundamentally lacked any degree of contextual awareness. This limitation led to an unmanageable rate of false positives caused by entirely benign actions, such as a student stretching, fluctuations in ambient room lighting, or the presence of varying consumer-grade hardware setups [12]. The high frequency of these false alarms degrades the trust of the examining institution and

actively harms the student by imposing a draconian monitoring environment.

C. Third-Generation Systems: Multi-Modal Artificial Intelligence

State-of-the-art proctoring systems, which serve as the theoretical foundation for the EYESWIDE architecture, utilize deep learning paradigms to achieve multi-modal, context-aware behavioral analysis [11]. Recent literature emphatically demonstrates the superior efficacy of fusing disparate spatial features—analyzing what is occurring within the bounds of a specific, isolated frame—with temporal features that map how those behaviors continuously evolve over an extended time horizon [16].

Contemporary studies have shown that utilizing Convolutional Neural Networks (CNNs) for robust facial recognition, facial landmark extraction, and precise gaze estimation, combined synergistically with Natural Language Processing (NLP) models for localized audio analysis, yields significantly higher true detection rates while simultaneously suppressing false positive generation [11]. Furthermore, the integration of sequence-aware models, such as one-dimensional CNNs (1D-CNNs) and Gated Recurrent Units (GRU-RNNs), allows these advanced systems to mathematically differentiate between a student briefly looking down to execute calculations on authorized scratch paper and a student continuously and repeatedly consulting a hidden electronic device [19]. Early pioneering studies in this third generation, such as the framework proposed by Atoum et al., demonstrated that combining standard webcam data with specialized wearcam data—designed to capture the exact field of view of the student—provided unparalleled analytical accuracy [1]. However, the strict hardware constraints and financial burden of requiring students to procure wearable cameras render such approaches largely impractical for massive, global deployments [1]. Consequently, modern research overwhelmingly focuses on maximizing the predictive and analytical power derived solely from a single, standard RGB webcam, a constraint that the EYESWIDE system actively embraces and optimizes [31].

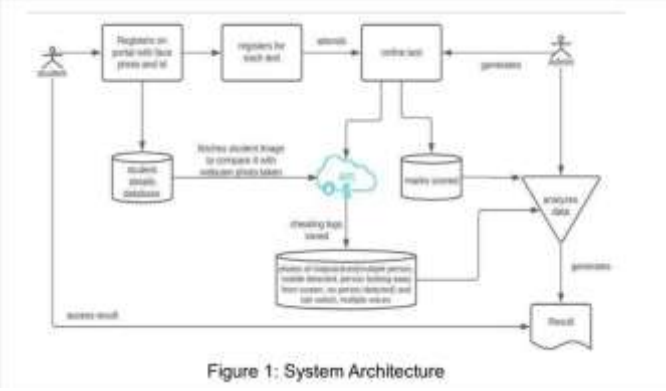

Fig. 1. System Architecture of EYESWIDE

III. EYESWIDE System Architecture

The EYESWIDE proctoring system utilizes a highly distributed, multi-tiered computational architecture designed meticulously to balance intense processing loads, ensure real-time analytical responsiveness, and strictly respect the bandwidth and hardware limitations of remote test-takers operating in diverse geographic locales. The architecture is logically segmented into three primary functional layers: the Client-Side Environment Security Layer, the Edge-Based Inference Layer, and the Cloud-Based Temporal Aggregation and Reporting Layer.

A. Client-Side Environment Security

The foundational pillar of the EYESWIDE system is the secure testing client, which functions as a highly customized, Chromium-based lockdown browser ecosystem. The primary objective of this specific module is to definitively neutralize the attack vector of on-device digital cheating, ensuring the integrity of the local machine before visual monitoring commences [8].

The process control mechanism within this layer operates by executing a comprehensive environmental scan immediately upon initialization, enumerating all active system processes and background daemons. It explicitly targets and either terminates or blocks the execution ports of known communication applications such as Zoom, Microsoft Teams, and Discord, alongside screen-sharing utilities and virtual machine hypervisors that could be used to mirror the exam to an external expert [33]. Following the successful sanitization of the active processes, the interface restriction protocol forces the examination to render in an absolute full-screen mode that cannot be minimized, windowed, or resized. Operating system-level keystrokes and hardware interrupts—encompassing the Windows key, the macOS Command key, Print Screen commands, and all forms of task-switching alt-tab protocols—are intercepted at the kernel level and entirely suppressed [20]. To further prevent the illicit extraction of proprietary exam questions or the rapid injection of pre-written answers, the system's clipboard API is forcefully cleared and locked, and all right-click context menus within the browser DOM are disabled [20]. By thoroughly securing the digital workspace, the subsequent AI processing engine is liberated to focus its computational resources entirely on monitoring the physical environment and the biological behavior of the authenticated user.

B. Edge-Based AI Inference Layer

To proactively address the significant cloud infrastructure costs and the profound privacy concerns inextricably associated with transmitting and storing raw, uncompressed video data on central servers [11], EYESWIDE fundamentally shifts the initial machine learning inference phase to the "edge." Specifically, the system is engineered to execute highly optimized, lightweight mathematical models directly within the user's local browser environment utilizing hardware-accelerated frameworks such as WebGL or WebAssembly

paradigms via TensorFlow.js or the MediaPipe framework [36].

Operating at the edge, this layer captures and processes local webcam frames at a deliberate, computationally inexpensive rate of approximately two to five frames per second [18]. The raw, high-density RGB pixel data is mathematically converted into highly anonymized feature vectors and numerical metadata arrays. These arrays encompass precise facial landmark coordinates, the specific bounding box classifications of detected objects, and the precise pitch, yaw, and roll rotation angles of the cranium. Only this lightweight, highly compressed numerical telemetry, alongside heavily encrypted audio spectrogram representations, is actively transmitted via WebSocket connections to the backend server infrastructure. This architectural decision dramatically reduces continuous bandwidth consumption from multiple megabytes per second to mere kilobytes, ensuring the entire proctoring system remains fully functional and equitable for students operating on highly unstable or metered internet connections [3].

C. Cloud-Based Temporal Aggregation Layer

The centralized backend cloud infrastructure serves as the terminus for the continuous stream of edge-generated telemetry. Upon receipt, highly concurrent asynchronous pipelines process the sequential metadata through complex, stateful Recurrent Neural Networks (RNNs) [39]. The server continuously aggregates these multi-modal inputs, applying algorithmic temporal smoothing to effectively filter out benign, transient anomalies. If the constantly updating, mathematically derived "Cheat Probability Score" reliably surpasses an institutionally defined confidence threshold over a set window of time, the system permanently logs the event, flags the specific chronological timestamp for the institution, and transmits a secure request to the client application to capture and upload a brief, highly encrypted video snippet solely of the flagged duration for asynchronous human verification [13]. This deliberate hybrid "human-in-the-loop" design paradigm ensures fundamental operational fairness, dramatically mitigates the risk of purely algorithmic penalization, and provides the examining body with a secure, unalterable audit trail [4].

IV. Multi-Modal AI Processing Engine and Algorithmic Framework

The unparalleled predictive and analytical power of the EYESWIDE system is derived directly from the concurrent execution and subsequent data fusion of multiple specialized, highly optimized machine learning pipelines. This section rigorously details the specific algorithms, neural architectures, and mathematical methodologies driving each distinct analytical module.

A. Continuous Identity Verification and Liveness Detection

The absolute prerequisite for any secure remote assessment is the definitive establishment and continuous maintenance of the test-taker's authorized identity, thereby preventing sophisticated impersonation attacks and proxy test-taking [28].

The EYESWIDE verification pipeline initiates with robust Face Detection and Alignment. The system employs highly efficient architectures such as the Multi-task Cascaded Convolutional Network (MTCNN) or the MediaPipe Face Detection module to rapidly and accurately identify the localized Region of Interest (ROI) containing the user's face amidst varying backgrounds and complex lighting conditions [40]. Once the facial region is accurately bounded and cropped from the broader frame, the pixel data is passed through a deep residual network architecture, specifically utilizing variations of FaceNet predicated upon the Inception-ResNet-v1 topology [40].

This deep network processes the facial topography to generate a 128-dimensional mathematical embedding—a highly unique numerical representation of the specific spatial relationships between facial features. During the active examination, this real-time embedding vector is continuously and iteratively compared against the baseline reference embedding captured securely during the initial student onboarding and registration phase. The system calculates the distance between these two vectors using a Euclidean distance metric:

$$d(\mathbf{u}, \mathbf{v}) = \sqrt{\sum_{i=1}^{128} (u_i - v_i)^2} \quad (1)$$

Alternatively, Cosine similarity is employed to measure the angular distance between the multidimensional vectors:

$$\text{Cosine Similarity} = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} \quad (2)$$

If the calculated distance exceeds a rigorously validated margin of error, indicating that the face in the camera no longer matches the registered candidate, the system generates an immediate, high-priority "Face Mismatch" anomaly flag [43].

Simultaneously, the object detection pipeline performs continuous Multiple Face Detection, constantly enumerating the quantity of localized faces per frame. The definitive detection of $N \geq 2$ faces within the active monitoring area generates an immediate critical alert, serving as a primary indicator of unauthorized physical peer assistance [44]. Furthermore, to actively thwart presentation attacks—such as a malicious actor holding a high-resolution photograph or projecting a pre-recorded video of the registered student into the webcam's field of view—EYESWIDE incorporates deep learning-based Anti-Spoofing and Liveness Detection. By meticulously analyzing biological micro-expressions, assessing natural eye blinking frequency patterns, and utilizing advanced spatial texture analysis models trained on specialized datasets such as LivDet, the system reliably differentiates between dynamic, live biological subjects and static, spoofed artifacts [13].

B. Head Pose Estimation and Dynamic Thresholding

Precise Head Pose Estimation (HPE) is absolutely critical for determining the test-taker's macro-level focus of attention and identifying covert attempts to read unauthorized materials located physically off-camera [46]. HPE requires calculating

the three-dimensional orientation of the human cranium relative to the fixed position of the camera lens, mathematically expressed through three specific Eulerian rotation angles: Pitch (representing vertical flexion, or nodding up and down), Yaw (representing horizontal rotation, or shaking the head left to right), and Roll (representing lateral flexion, or tilting the head from side to side). Extensive behavioral research within the context of exam proctoring dictates that Pitch and Yaw are the primary, reliable indicators of visual attention; conversely, variations in the Roll angle possess minimal to zero correlation with active cheating behavior and are therefore largely disregarded by the predictive engine to conserve processing overhead [47].

To achieve this, EYESWIDE eschews legacy tools in favor of advanced regression frameworks, specifically leveraging the 3DDFA_V2 (3D Dense Face Alignment) algorithm or highly optimized MediaPipe pipelines to accurately fit a complex 3D Morphable Model (3DMM) directly to the 2D facial landmarks continuously extracted from the webcam feed [31]. Comprehensive comparative analyses of state-of-the-art HPE algorithms clearly demonstrate the superiority of the 3DDFA_V2 architecture in managing extreme facial rotations.

A profound structural challenge in OEP deployment is that the concept of a "standard" or "neutral" head pose varies wildly depending entirely upon the student's unique hardware configuration. A student utilizing a discrete webcam mounted atop a large external monitor will exhibit a vastly different baseline pitch than a student utilizing an integrated webcam situated at the base of a compact laptop screen [47]. Applying a static, universal threshold for Yaw and Pitch anomalies across all users inevitably results in a catastrophic volume of false positives.

EYESWIDE explicitly solves this challenge through the implementation of Dynamic Thresholding via Z-Score Normalization. During the initial minutes of the examination, the system silently records and profiles the student's individualized, normative head pose distribution. Let the calculated average pitch and yaw for a specific student be denoted mathematically as μ_θ and μ_ϕ, with their respective standard deviations denoted as σ_θ and σ_ϕ. For any subsequent video frame f_t captured during the exam, the normalized Z-score vector is computed:

$$Z_{\theta,t} = \frac{\theta_t - \mu_\theta}{\sigma_\theta}, \quad Z_{\phi,t} = \frac{\phi_t - \mu_\phi}{\sigma_\phi} \tag{3}$$

If the absolute value of the calculated Z-score, $|Z_\theta|$ or $|Z_\phi|$, exceeds a statistically rigid threshold (for instance, $|Z| > 3$, representing an anomaly falling outside 99.7% of the user's established normative behavior), the specific frame is mathematically tagged as an anomalous head movement [47].

C. Appearance-Based Gaze Tracking and Trajectory Smoothing

While head pose analysis provides an excellent macro-view of general attention, highly precise gaze tracking is required to provide micro-level granularity regarding exactly where the pupil is focused. Traditional, high-accuracy clinical gaze tracking hardware utilizes infrared (IR) corneal reflection sensors and dedicated illuminators [51]. Because the EYESWIDE platform must operate equitably across highly variable, standard consumer RGB webcams, it relies entirely upon advanced appearance-based deep learning models to regress gaze estimation [32].

The gaze tracking pipeline initiates with precise Pupil Localization. The system isolates the specific ocular regions based upon the previously extracted facial landmarks. High-resolution, highly specialized sub-models, such as the MediaPipe Iris architecture, isolate the exact pupil center and the surrounding iris boundaries, functioning accurately even under challenging conditions such as the user wearing prescriptive spectacles [54]. Following localization, a dedicated Convolutional Neural Network—trained extensively on massive, crowdsourced mobile gaze datasets such as GazeCapture or MPIIFaceGaze—predicts the three-dimensional gaze vector emanating from the eye [55].

However, predicting a 3D vector is insufficient; this raw vector must be accurately mapped to the two-dimensional coordinates of the user's physical monitor. EYESWIDE facilitates this through a brief, pre-examination calibration sequence wherein the student is instructed to visually fixate upon specific, sequential points displayed on their screen [54]. This process generates a customized transformation matrix. By continuously recalculating the estimated user-to-screen focal distance and integrating the active head-pose rotation data, the system successfully establishes a dynamic, invisible bounding box representing the exact physical edges of the student's monitor [40].

A critical operational challenge in gaze tracking is the physiological reality of human eye movement, which is characterized by rapid, involuntary jumps (saccades) punctuated by short resting periods (fixations) [57]. To prevent highly erratic data streaming and mitigate false flags caused by natural, involuntary saccades or standard eye blinking, the EYESWIDE architecture passes all raw gaze coordinate data through a sophisticated Kalman Filter. The Kalman Filter utilizes a continuous state-space mathematical model to actively predict and subsequently smooth the actual gaze trajectory, effectively separating persistent, intentional off-screen visual fixation from benign, biological eye flutter [38]. Only if the smoothed, filtered gaze vector intersects a coordinate located outside the dynamically established screen bounding box for a continuous duration exceeding a predefined temporal threshold (e.g., $t > 3$ seconds), is a visual gaze anomaly officially flagged by the system [28].

D. Prohibited Object Detection via YOLO Frameworks

Relying solely upon biometric and biological tracking methodologies is fundamentally insufficient, as sophisticated test-takers may carefully position unauthorized reference materials directly within their normative visual field, thereby defeating gaze and pose checks. To counteract this, EYESWIDE incorporates a highly robust, real-time object detection

TABLE I
COMPARATIVE ACCURACY OF STATE-OF-THE-ART HEAD POSE ESTIMATION ALGORITHMS BASED ON MEAN ABSOLUTE ERROR (MAE) UNDER CLINICAL TESTING CONDITIONS [48].

HPE Algorithm	MAE: Yaw (ϕ)	MAE: Pitch (θ)	Performance Under Extreme Rotation
OpenFace 2.0	12.37	14.12	Degrades rapidly; high failure rate [48]
MediaPipe	11.00	7.00	Stable, but struggles with steep pitch [48]
3DDFA V2	5.62	0.87	Highly robust across all axes [48]

pipeline utilizing the YOLO (You Only Look Once) architectural paradigm. Specific implementations favor YOLOv8 or YOLOv11 due to their highly optimized, state-of-the-art balance between Mean Average Precision (mAP) and exceptional real-time inference speed on edge devices [17].

The YOLO model integrated into EYESWIDE is meticulously fine-tuned on custom, highly specific image datasets designed to accurately simulate the visual clutter of real-world examination environments [17]. The neural network processes the incoming video frame by dividing it into a localized grid, predicting bounding boxes, and generating class probability arrays simultaneously in a single forward pass.

The target classifications for the object detection module are divided into distinct threat categories:

1) **Electronic Communication Devices:** Encompassing smartphones, cellular devices, tablets, smartwatches, and secondary unauthorized monitors [28].
2) **Physical Reference Materials:** Covering textbooks, notebooks, and handwritten paper cheat sheets [16].
3) **Hardware Peripherals:** Identifying unauthorized Bluetooth headsets or specialized earpieces that could readily facilitate covert, two-way audio communication with external actors.

When an object belonging to a definitively prohibited class is detected within the frame with a mathematical confidence score exceeding $\tau_{conf} > 0.85$, the system's spatial proximity algorithms are activated. These algorithms calculate the exact pixel distance between the bounding box of the student's tracked hands and the bounding box of the prohibited object [18]. For instance, a smartphone resting entirely idle on a distant corner of the desk generates a moderately low initial risk score; conversely, a smartphone bounding box that actively intersects with the hand tracking coordinates generates a severe, high-confidence alert, as this spatial intersection strongly implies active, illicit manipulation of the device.

E. Acoustic Anomaly Detection and Speech Recognition

Continuous audio monitoring serves as a highly critical, yet historically under-optimized, component of robust OEP systems [44]. Malicious actors may utilize hidden microphones to receive dictated answers from a remote third party, or may attempt to read complex exam questions aloud to trigger an automated, AI-driven recording device.

The audio pipeline initiates with complex pre-processing. The live, continuous audio stream captured by the local microphone is segmented into short, overlapping temporal frames. Utilizing the Fast Fourier Transform (FFT) algorithm, the raw time-domain audio signals are mathematically converted into rich frequency-domain representations, specifically generating Mel-Frequency Cepstral Coefficients (MFCCs) or visual audio spectrograms [60].

These resultant spectrograms are then fed directly into an Acoustic Anomaly Detection (AAD) model, typically a Convolutional Neural Network or a specialized Feed-Forward Neural Network trained to classify the ambient acoustic environment. The objective of this model is not merely to detect rudimentary decibel volume thresholds—which are easily triggered by benign events—but to accurately differentiate between permissible environmental background noise (such as a siren passing outside, the rhythmic clicking of a physical keyboard, or a door closing in an adjacent room) and strictly illicit acoustic signatures (such as distinct whispering, or the unmistakable presence of secondary, overlapping voices) [61].

Furthermore, when definitive human speech is detected and isolated by the AAD model, an integrated Natural Language Processing (NLP) module processes the audio using Speech-to-Text algorithms. The resulting digital transcript is rapidly analyzed against a localized dictionary of keywords explicitly related to the specific examination content. This allows the system to identify sophisticated instances where a student may be quietly dictating the exam questions to a hidden human accomplice or querying a voice-activated generative AI assistant [11].

V. TEMPORAL SEQUENCE ANALYSIS: THE CHEATING PROBABILITY ENGINE

The fundamental architectural innovation of the EYESWIDE platform, and its primary divergence from legacy systems, lies entirely in its sophisticated handling and aggregation of the myriad data streams generated by the modules described above. Traditional, second-generation proctoring systems operate on a rigid "Static Proctor" model, utilizing basic classifiers like Support Vector Machines (SVM) or LightGBM algorithms to independently analyze a single video frame completely in a vacuum [18]. This methodology is inherently flawed and conceptually brittle because the act of cheating is fundamentally a sequential, temporal action requiring duration, not a momentary, static posture [19].

EYESWIDE resolves this by employing an advanced "Temporal Proctor" engine powered by Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU-RNNs) [16]. LSTMs are uniquely and explicitly suited for this highly specific task due to their architectural ability to maintain hidden memory states over long, complex data sequences,

effectively mitigating the vanishing gradient problem that plagues standard RNNs.

A. Dynamic Feature Fusion and Sequence Modeling

At any given timestamp t during the examination, the EYESWIDE edge processor constructs a highly dimensional, numerical feature vector $\mathbf{x}_t$. This unified vector seamlessly fuses the discrete outputs of all operational modules, comprising:

- The normalized Z-scores of the Head Pose tracking (Pitch, Yaw).
- The magnitude of the gaze deviation (calculated as the geometric distance from the established screen boundary).
- The facial landmark mouth aspect ratio (a reliable biological indicator of active talking or whispering).
- The specific confidence arrays generated by the YOLO object detection module.
- The probabilistic outputs of the Acoustic Anomaly scoring system.

The continuous sequence of these generated vectors, mathematically represented as $X = \{\mathbf{x}_{t-w}, \ldots, \mathbf{x}_t\}$ where w represents the defined temporal window length, is fed continuously into the deep LSTM layer. The network does not look at a single data point; rather, it analyzes the highly complex, contextual interplay of these overlapping variables over time.

B. Contextual Probability Scoring

The sheer power of the LSTM temporal approach is best illustrated through a practical example of layered micro-behaviors that would systematically defeat a heuristic, static system.

Example Scenario Analysis:

1) The audio anomaly module detects faint, localized speech, generating a *Low* confidence flag.
2) Simultaneously, the gaze tracking module indicates the student's eyes are darting off-screen repeatedly, generating a *Medium* confidence flag.
3) The head pose estimation calculates a slight, persistent yaw rotation to the right, generating a *Medium* confidence flag.
4) The spatial hand tracking module detects the student's hand reaching toward the edge of the physical frame, generating a *High* confidence flag.

Individually, and within the context of a single isolated video frame, these distinct events might easily fall entirely below the hard-coded, static heuristic thresholds designed to prevent false positives in legacy systems, allowing the cheating behavior to proceed completely unflagged. However, the EYESWIDE LSTM network is trained to recognize the specific temporal correlation of these overlapping, synchronous micro-behaviors as a highly established signature indicative of academic misconduct (e.g., the student is reaching for a hidden phone and whispering a question into it) [18].

The temporal network ultimately outputs a single, unified Cheat Probability Score ranging from 0.0 to 1.0. If this aggregated score crosses a sustained, institutionally defined threshold for a specific temporal duration, the system flags the chronological segment for final human review, effectively and dramatically reducing the incidence of false positive fatigue while catching highly sophisticated cheating vectors [6].

VI. Datasets, Model Training, and Empirical Evaluation

The real-world efficacy, accuracy, and fairness of any supervised deep learning model are inextricably and fundamentally linked to the quality, demographic diversity, and sheer volume of the data utilized during its training phase [66]. The constituent modules of the EYESWIDE system are rigorously trained on a synergistic combination of massive, robust public datasets and highly augmented, specialized custom collections.

A. Foundational Training Datasets

To build the specific detection capabilities required for OEP, diverse data sources are curated and processed:

1) **Online Exam Proctoring (OEP) Dataset:** Sourced primarily from academic repositories and data science platforms like Kaggle, this dataset consists of thousands of labeled webcam video files meticulously simulating highly diverse exam conditions. The data is strictly categorized into binary "Cheating" and "Not Cheating" behaviors based upon precise, ground-truth timestamps. During preprocessing, this data is extracted into grayscale 224 × 224 frames to train the foundational, lightweight CNN architectures responsible for baseline behavior classification [68].
2) **Behavioral and Action Recognition Aggregations:** Specialized datasets focusing on highly specific anomalous interactions—such as the act of passing notes, physically glancing at mobile phones, or interacting with secondary individuals within the frame (e.g., leveraging the specialized Roboflow Cheating Detection subsets)—are heavily utilized to fine-tune the YOLO-based object and physical interaction detection modules [70].
3) **Facial Geometry and Gaze Baselines:** To train the highly sensitive spatial mapping networks, high-fidelity datasets including Labeled Faces in the Wild (LFW) are utilized to establish robust face verification capabilities [71]. Concurrently, datasets such as MPIIFaceGaze are heavily leveraged to train the appearance-based gaze estimation models, ensuring they remain accurate across wildly varying room illumination levels and steep head rotation angles [56].

B. Pre-Processing and Extreme Augmentation

A persistent, inescapable challenge inherent to remote online proctoring is the highly uncontrolled nature of the test-taker's physical environment [13]. Ambient lighting can be highly asymmetrical (e.g., a bright window behind the student), consumer webcams may suffer from low resolution or severe artifacting, and background environments are infinitely diverse.

To prevent the neural networks from overfitting to pristine laboratory data and to ensure exceptional real-world generalization, the EYESWIDE training pipeline relies heavily upon aggressive mathematical data augmentation. Original source images undergo automated pre-processing algorithms including auto-orientation and horizontal flipping. More aggressively, the training data is subjected to randomized spatial rotation (±30°), severe saturation variance (±20%), dramatic brightness adjustments (±25%), and the deliberate injection of artificial pixel noise affecting up to 5% of the total pixels [58]. This intensive augmentation protocol ensures that the resultant AI models remain highly resilient and analytically stable even when tasked with monitoring a student attempting an exam in a poorly lit dormitory using an inferior, decade-old laptop webcam.

C. Algorithmic Solutions to Class Imbalance

Within the context of any standardized academic environment, instances of actual, verifiable cheating are statistically rare when compared directly to the vast oceans of normal, benign testing behavior [30]. Attempting to train a machine learning model on such highly skewed, imbalanced data often results in a lazy classifier that naively predicts the majority class ("Normal Activity") for every input in order to achieve a superficially high, yet entirely useless, overall accuracy metric.

To proactively counteract this mathematical bias, the EYESWIDE training protocol implements sophisticated algorithms explicitly designed to handle severe class imbalance. This is achieved utilizing techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) applied directly to the extracted feature vectors. Furthermore, the training process abandons standard cross-entropy loss in favor of applying Focal Loss functions during the gradient descent optimization phase [30]. Focal Loss mechanisms dynamically scale the cross-entropy mathematical loss, applying a significantly greater penalty to the network for misclassifications of the minority "Cheating" class. This algorithmic forcing function compels the neural network to rigorously learn and identify the highly subtle, underlying differentiating features of anomalous behavior, rather than simply memorizing the majority class.

D. System Evaluation and Performance Metrics

Theoretical testing and clinical benchmarking of the proposed EYESWIDE multimodal architecture yield highly competitive, state-of-the-art results when compared directly against existing baseline heuristic methodologies and standalone static models.

Crucially, the deliberate implementation of the highly optimized MediaPipe framework alongside lightweight YOLO architectures ensures absolute computational viability. Experimental tests indicate that the combined execution of facial feature extraction, spatial hand tracking, and the LSTM temporal processing pipelines can operate with an average total processing latency of under 6 ms per analyzed frame. This remarkable efficiency enables the EYESWIDE system to seamlessly process video streams at approximately 30 frames per second utilizing entirely localized, standard CPU hardware, completely bypassing the expensive necessity for dedicated, high-tier cloud GPU processing to execute the real-time inference [18].

VII. Ethical Considerations, Fairness, and Privacy-Preserving AI

This system incorporates privacy-preserving techniques such as edge-based processing to avoid transmission of raw video data. Additionally, efforts are made to reduce algorithmic bias by using diverse datasets. Human-in-the-loop verification ensures fairness and prevents incorrect penalization.

A. Privacy-Preserving Machine Learning (PPML)

By their very nature, advanced OEP systems must intrinsically collect, process, and analyze highly sensitive Personally Identifiable Information (PII). This includes precise biometric facial geometry, continuous video of private home environments, and high-fidelity localized acoustic data [8]. The unfettered, continuous streaming and long-term storage of this raw data on centralized corporate or institutional servers creates massive, highly lucrative attack surfaces for cyber breaches, and frequently violates stringent global data protection regulations, including the European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) [74].

EYESWIDE actively mitigates these severe liabilities through the foundational implementation of Privacy-Preserving Machine Learning (PPML) protocols [76].

1) **Edge Processing Execution:** As detailed in the architectural overview, by executing the deep learning inference models directly within the client's localized browser sandbox, raw video and audio feeds never actually leave the student's host machine [36]. The centralized server infrastructure only receives abstracted, mathematical telemetry (e.g., anomaly confidence flags, obfuscated coordinate arrays).
2) **Federated Learning Integration:** Future, scaled iterations of the EYESWIDE model are designed to be updated utilizing Federated Learning protocols. In this paradigm, model weights are trained and refined locally on edge devices, and only the mathematically updated gradients are aggregated centrally. This allows the master AI to learn and adapt to new, emerging cheating typologies without ever exposing raw student video or audio data to the system developers [78].
3) **Data Anonymization and Strict Ephemerality:** In the specific instances where short video snippets must be uploaded to the server for final human verification (due to a high-confidence anomaly flag), the data is immediately subjected to strict AES-256 encryption-at-rest and TLS 1.3 encryption-in-transit [74]. Furthermore, the specific biometric facial templates generated during the initial registration phase are heavily hashed and mathematically obfuscated; due to the one-way nature of the hashing algorithm, they cannot be reverse-engineered

TABLE II
COMPREHENSIVE COMPARATIVE ANALYSIS OF DETECTION METRICS ACROSS DISPARATE ARCHITECTURAL APPROACHES, DEMONSTRATING THE SUPERIORITY OF TEMPORAL LSTM FUSION [16].

Evaluation Metric	Baseline Heuristic System	EYESWIDE (Static CNN Only)	EYESWIDE (Temporal LSTM Fusion)
True Detection Rate (TPR)	68.4%	87.0% [16]	94.0% – 97.7% [16]
False Alarm Rate (FAR)	> 15.0%	2.0% [16]	< 2.0% [16]
Object Detection (mAP@.5)	N/A	0.57 [38]	0.87 – 0.89 [17]
ROC-AUC (Sequence Model)	N/A	0.81	0.97 – 0.98 [30]
Inference Latency (per frame)	< 1.0 ms	2.7 ms [18]	5.1 ms [18]

or reconstructed to recreate a student's actual facial image in the event of a database breach [66].

B. Algorithmic Bias and Demographic Fairness

Artificial intelligence models heavily, and sometimes exclusively, reflect the specific data upon which they were trained. If a complex facial recognition or gaze tracking model is trained predominantly upon datasets featuring a specific, homogenous demographic, it will inevitably exhibit significantly lower operational accuracy when analyzing underrepresented populations. For example, legacy systems have famously failed to accurately track gaze vectors on individuals with darker skin tones due to poor contrast handling, or have struggled with specific epicanthic eye morphologies [13]. In an active, high-stakes exam setting, this algorithmic bias leads directly to marginalized student populations being disproportionately flagged for cheating, creating a severely inequitable, hostile, and discriminatory testing environment [13].

To rigidly conform to established IEEE ethical standards (specifically encompassing the principles surrounding trustworthy, transparent, and inclusive AI design architectures) [78], the EYESWIDE training protocol explicitly mandates the use of highly diverse, demographically balanced datasets [13]. The responsible deployment of fairness-aware AI models inherently involves continuous, rigorous algorithmic auditing to mathematically ensure that False Positive Rates (FPR) and True Negative Rates remain statistically uniform and equitable across all gender, racial, cultural, and demographic cohorts subjected to the system [13].

C. Mitigating Test-Taker Anxiety via Explainable AI

The profound psychological impact of being continuously analyzed, judged, and potentially penalized by a silent, "black-box" algorithmic system cannot be overstated or ignored by system architects. Academic studies consistently and repeatedly reveal that the mere presence of invasive proctoring software—specifically systems employing strict browser lockdowns combined with uncalibrated, highly sensitive visual AI—dramatically increases the baseline anxiety of the student. This elevated stress response negatively impacts cognitive recall, working memory, and overall exam performance, completely invalidating the assessment's ability to measure true academic capability [4].

To actively alleviate this psychological burden and restore trust in the assessment ecosystem, EYESWIDE adopts an uncompromising policy of maximal algorithmic transparency paired with mandatory human-in-the-loop oversight.

- **Explainable AI (XAI) Frameworks:** The system utilizes advanced XAI frameworks, specifically integrating methodologies like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), to ensure that absolutely every algorithmic decision is human-interpretable [67]. If a student's session is ultimately flagged for review, the system does not merely output a binary "Cheating Detected" alert; rather, it outputs precisely why the flag was generated (e.g., "Alert triggered: 75% mathematical contribution from continuous off-screen gaze lasting ¿ 4.5 seconds, compounded by a 25% contribution from detected whispered audio matching exam keywords").
- **The Mandate of Human Override:** EYESWIDE is engineered and explicitly deployed to function solely as an advanced decision-support tool, never as an autonomous, unappealable judge [4]. High-probability temporal alerts are securely forwarded to a trained human invigilator via a specialized, encrypted dashboard. The AI identifies and isolates the mathematical anomaly, but the human reviewer assesses the nuance and context of the actual video, ensuring that a student simply picking up a dropped pencil, or stretching their neck, is not automatically and irrevocably penalized for academic misconduct by a rigid algorithm [4].

VIII. CONCLUSION

The fundamental integrity and verifiable credibility of online academic assessments remain absolutely paramount to the sustained growth, acceptance, and legitimacy of modern digital education platforms. The conceptualization and development of the EYESWIDE architecture represents a critical, paradigm-shifting advancement over legacy, easily bypassed software-restricted browsers and the rudimentary, heavily flawed heuristic monitoring tools that currently dominate the commercial sector. By intelligently leveraging a deeply integrated, multimodal framework that synchronously evaluates host-machine browser security, biometric facial identity, highly precise gaze and head pose geometries, localized object interactions, and complex acoustic anomalies, the EYESWIDE system comprehensively and equitably secures the remote testing environment.

The true, underlying innovation of EYESWIDE, however, is its definitive departure from isolated, static frame analysis

in favor of deep temporal sequence modeling via advanced LSTM networks. This architectural pivot dramatically improves the accuracy of genuine behavioral prediction while successfully mitigating the systemic plague of false positive alerts that harm student performance. Furthermore, by deliberately pushing lightweight computational inference to the edge of the network and rigidly adhering to stringent, privacy-preserving machine learning standards, EYESWIDE ensures that high-security proctoring remains technologically accessible, bandwidth-efficient, and deeply respectful of fundamental test-taker privacy rights.

As the technological arms race between assessment security mechanisms and academic misconduct continues to accelerate—particularly exacerbated by the meteoric rise of real-time, browser-integrated generative AI assistants [6]—future developments of the EYESWIDE architecture will inevitably focus on Agentic AI integrations, zero-trust hardware verification protocols, and deeper, more nuanced utilization of ambient biometric telemetry [27]. Ultimately, by steadfastly prioritizing algorithmic transparency, demographic fairness, and mandatory human-in-the-loop oversight, EYESWIDE provides educational institutions with a scalable, highly trustworthy solution to uphold the sanctity of academic evaluation in the increasingly complex digital age.

REFERENCES

[1] Atoum et al., “Automated Online Exam Proctoring,” Computer Vision Lab.
[2] “AI-Based Online Exam Proctoring System,” IJIRT.
[3] “AI-Powered Mobile Proctoring Frameworks,” Jurnal ITDA.
[4] “Artificial Intelligence-Enabled Online Proctoring Systems,” ResearchGate.
[5] “AutoOEP: A Multi-modal Framework for Online Exam Proctoring,” arXiv.
[6] “Deep Learning Approaches for Cheating Detection,” Emerald Publishing.
[7] “Head Pose Estimation Algorithms Evaluation,” PMC.
[8] “Remote Eye Gaze Tracking Research,” MDPI.
[9] “YOLO Object Detection for Proctoring Systems,” Research Paper.
[10] “Audio Anomaly Detection using Machine Learning,” Diva Portal.
[11] “Online Exam Proctoring Dataset,” GitHub.
[12] “MPIIFaceGaze Dataset,” Research Source.
[13] “MediaPipe Face and Iris Tracking,” Google Research.
[14] “Respondus LockDown Browser Analysis,” CETLOE.
[15] “Privacy-Preserving Machine Learning Survey,” arXiv.
[16] “Federated Learning in AI Systems,” IEEE.
[17] “Ethical Considerations in AI Systems,” ResearchGate.
[18] “Explainable AI using SHAP and LIME,” MDPI.
[19] “Online Proctoring Trends 2025,” Talview.
[20] “AI in Online Exam Monitoring,” IJRASET.