DeepLens Engine for Focus: A Multi-Layered Framework for Real-Time Attention Monitoring in Productivity Environments

Nafis Aslam
School of Computer Sciences
Universiti Sains Malaysia
Penang, Malaysia
nafisaslam18@student.usm.my

Abstract—The proliferation of digital distractions has made sustained attention increasingly difficult to maintain. Studies indicate that average screen attention spans have declined from 2.5 minutes in 2004 to approximately 47 seconds in recent years, with interruption recovery requiring over 20 minutes on average. While artificial intelligence-based attention monitoring systems have emerged as potential solutions, existing approaches suffer from three fundamental limitations: lack of standardized attention state definitions, reliance on single-modal detection methods, and absence of temporal context leading to excessive false positives.

This paper presents the **DeepLens Engine for Focus (DLEF)**, a novel multi-layered framework addressing these limitations through systematic integration of deep learning classification, auxiliary validation, and temporal heuristic smoothing. Grounded in productivity science—synthesizing insights from deep work theory, deliberate practice research, flow psychology, and execution frameworks—we establish formal definitions for six attention states: Focused, Phone, Absent, Drowsy, Looking Away, and Bad Posture, operationalized through observable physical indicators.

The DLEF architecture comprises four synergistic layers: (1) a primary YOLOv11n-cls classifier trained on approximately 3,000 augmented images derived from 1,080 base samples, (2) auxiliary validation modules utilizing MediaPipe for gaze estimation and eye aspect ratio computation, (3) a temporal heuristic layer implementing the novel 20/30-second rule for sustained distraction confirmation, and (4) a context-aware response layer for non-intrusive intervention.

Experimental evaluation on 210 held-out validation images demonstrates 94.76% classification accuracy, with 100% precision on critical classes including Phone and Absent. The temporal heuristic layer substantially reduces false positives compared to frame-by-frame classification while maintaining high sensitivity. Real-time performance achieves 15–18 FPS on consumer hardware. We implement the complete framework as **DeepWork AI**, a privacy-preserving web application with Pomodoro-style session management and comprehensive analytics.

Index Terms—Attention monitoring, focus detection, computer vision, convolutional neural networks, YOLO, real-time classification, human-computer interaction, productivity systems, deep learning, temporal heuristics.

1 Introduction

The contemporary digital landscape presents unprecedented challenges to human attention. The average knowledge worker experiences interruptions approximately every 11 minutes [1], with research demonstrating that recovery to the original task requires an average of 23 minutes and 15 seconds [2]. More alarming, longitudinal studies document a troubling trend: attention spans on digital screens have declined from approximately 2.5 minutes in 2004 to merely 47 seconds by 2023 [3]—representing an 82% reduction in sustained attention capacity over less than two decades.

The economic and cognitive implications of this attentional degradation are substantial. Newport introduced the concept of "deep work"—professional activities performed in distraction-free concentration that push cognitive capabilities to their limit—arguing that such focused work is becoming simultaneously more valuable and more rare [4]. Research demonstrates that multitasking reduces productivity by up to 40% [5], while context switching carries significant cognitive overhead through "attention residue," where part of attention remains on the previous task even after switching [6].

While various software-based productivity tools exist—including website blockers, time trackers, and notification managers—these solutions primarily address *digital* distractions and rely on self-reporting or indirect metrics. They fundamentally fail to capture *physical* distractions such as phone checking, leaving the workspace, or drowsiness. Moreover, manual time tracking is notoriously unreliable, with users frequently forgetting to start or stop timers.

1.1 Design Philosophy

The design of DLEF is grounded not merely in computer vision capabilities, but in a synthesis of established productivity and performance science literature. Newport's concept of deep work [4]—cognitively demanding tasks performed in distraction-free concentration—provides the foundational philosophy. This is complemented by Ericsson's research on deliberate practice [25], which demonstrates that focused, feedback-rich training drives expertise development. Csikszentmihalyi's flow theory [11] establishes that optimal performance emerges from sustained, immersive engagement. Coyle's investigation of talent hotbeds [26] reveals that deep practice—characterized by operating at the edge of ability with immediate error correction—accelerates skill acquisition. Finally, McChesney *et al.*'s 4 Disciplines of Execution (4DX) framework [27] provides an operational structure: focus on wildly important goals, act on lead measures, maintain a compelling scoreboard, and create a cadence of accountability.

A common thread emerges across these works: sustained, distraction-free focus paired with meaningful feedback enables exceptional outcomes. Yet existing productivity tools fail to operationalize this insight—they track time spent, not attention quality. The DeepWork AI application directly implements this theoretical synthesis: goal management embodies 4DX's focus on what matters most; Pomodoro-style sessions structure deliberate practice periods; real-time focus monitoring via DLEF provides the immediate feedback loop that Ericsson identifies as essential; and analytics create the accountability scoreboard.

1.2 Problem Statement

Advances in computer vision and deep learning have enabled AI-based attention monitoring systems deployed in driver monitoring [7], classroom engagement tracking [8], and e-learning platforms [9]. However, existing approaches exhibit three fundamental limitations:

Limitation 1: Definitional Ambiguity. There exists no standardized, computationally tractable definition of "focused" versus "distracted" states. Different systems employ varying criteria—some consider looking away for more than 2 seconds as distraction, others use composite engagement scores. This heterogeneity makes cross-system comparison difficult.

Limitation 2: Single-Modal Detection. Most existing systems rely on a single modality for attention classification—typically either gaze tracking, posture analysis, or facial expression recognition. Gaze-only approaches cannot detect internal mind-wandering when eyes remain on screen [10]. Posture-only methods may misclassify unique but focused working positions.

Limitation 3: Temporal Naïveté. Frame-by-frame classification, while technically accurate on individual images, produces noisy outputs when applied to continuous video streams. A brief glance away, a momentary eye closure, or a quick posture adjustment can all trigger false "distracted" classifications, leading to user frustration and reduced trust.

1.3 Research Contributions

This paper addresses these limitations through the following contributions:

1) **Formal Attention State Definitions:** We establish rigorous, measurable criteria for focused and distracted states based on physical and behavioral indicators, operationalized for computer vision detection through six distinct classes.
2) **Multi-Layered Detection Architecture:** We introduce the DeepLens Engine for Focus (DLEF), a novel framework integrating: (a) YOLOv11n-cls deep learning classification, (b) MediaPipe-based auxiliary validation for gaze and drowsiness detection, (c) temporal heuristic smoothing via the 20/30 decision rule, and (d) context-aware intervention mechanisms.
3) **Temporal Heuristic Innovation:** We propose the 20/30 rule—a temporal smoothing mechanism requiring that more than 20 of 30 consecutive seconds be classified as distracted before confirming a distraction event. This approach substantially reduces false positives while maintaining high sensitivity to genuine distractions.
4) **Practical System Implementation:** We demonstrate the framework's integration within DeepWork AI, a complete productivity application featuring goal management, Pomodoro-style sessions, and comprehensive analytics, while preserving privacy through on-device processing.
5) **Empirical Validation:** We provide experimental evaluation achieving 94.76% classification accuracy with 100% precision on critical classes, with transparent discussion of evaluation scope and limitations.

2 Related Work

2.1 Computer Vision Approaches to Attention Detection

2.1.1 Gaze Tracking and Eye Movement Analysis

Eye tracking has been the predominant modality for attention detection, based on the established link between gaze direction and visual attention [12]. Key metrics include fixation duration (longer fixations indicate deeper processing [13]), saccade frequency, and gaze dispersion.

However, gaze-based approaches have fundamental limitations. The "looking but not seeing" phenomenon demonstrates that visual and cognitive attention can dissociate—users may maintain appropriate gaze while experiencing mind-wandering [14]. Studies estimate that mind-wandering occurs 25–50% of waking hours [15], representing a significant blind spot for gaze-only systems.

2.1.2 Facial Expression and Drowsiness Detection

Facial expression analysis provides complementary information about cognitive and affective states. The Facial Action Coding System (FACS) [16] provides a systematic taxonomy of facial movements widely adopted in computer vision. The Eye Aspect Ratio (EAR) metric introduced by Soukupová and Čech [17] enables real-time drowsiness detection through eye landmark analysis.

2.1.3 Posture and Body Language Analysis

Body posture provides contextual engagement information. Upright, forward-leaning postures typically correlate with engagement, while slouching suggests disengagement [18]. MediaPipe Pose [19] enables real-time body keypoint extraction. Bosch *et al.* demonstrated that combining facial features with body posture improved engagement classification by 12% compared to facial features alone [20].

2.2 Deep Learning for Classification

The YOLO (You Only Look Once) family has evolved significantly since its introduction [21]. YOLOv5 introduced improved training techniques [22]. YOLOv8 unified detection, segmentation, and classification [23]. YOLOv11 represents the latest iteration with improved accuracy through enhanced feature extraction.

Trabelsi *et al.* achieved 76–85% accuracy detecting attentive versus inattentive students using YOLOv5 [8], demonstrating YOLO's applicability to attention monitoring. However, single-model approaches struggle with contextual understanding and temporal consistency.

2.3 Gaps in Current Methods

Despite significant progress, several gaps persist: (1) inconsistent definitions across studies, (2) single-modal limitations, (3) lack of temporal context producing noisy outputs, and (4) limited generalization from specific populations. Our work addresses these through formalized definitions, multi-modal fusion, temporal heuristics, and transparent reporting of evaluation scope.

3 Formal Definitions of Attention States

A core contribution of this work is formalizing "focused" and "distracted" states in terms rooted in cognitive theory and detectable via computer vision.

3.1 Attention State Space

We define the attention state space $\mathcal{S}$ as six mutually exclusive states based on observable indicators:

$$\mathcal{S} = \{S_F, S_{Ph}, S_{Ab}, S_{Dr}, S_{LA}, S_{BP}\} \tag{1}$$

TABLE 1: Attention State Definitions

State	Symbol	Observable Criteria
Focused	S_F	Eyes on task, upright posture, engaged expression
Phone	S_{Ph}	Interacting with mobile device
Absent	S_{Ab}	Not present in camera frame
Drowsy	S_{Dr}	EAR < 0.25, signs of fatigue
Looking Away	S_{LA}	Gaze diverted > 45 from screen
Bad Posture	S_{BP}	Slouched, disengaged position while present

3.2 Productive vs. Distracted Partition

We partition the state space into productive ($\mathcal{S}_p$) and distracted ($\mathcal{S}_d$) subsets:

$$\mathcal{S}_p = \{S_F\} \tag{2}$$

$$\mathcal{S}_d = \{S_{Ph}, S_{Ab}, S_{Dr}, S_{LA}, S_{BP}\} \tag{3}$$

The binary focus indicator at time t is:

$$y_t = \begin{cases} 1 & \text{if } s_t \in \mathcal{S}_p \\ 0 & \text{if } s_t \in \mathcal{S}_d \end{cases} \tag{4}$$

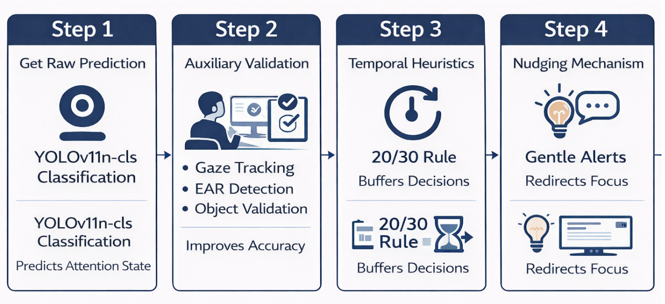


Fig. 1: The DLEF four-layer architecture. Layer 1 performs primary classification using YOLOv11n-cls. Layer 2 provides auxiliary validation through gaze estimation and EAR analysis. Layer 3 applies the 20/30 temporal heuristic. Layer 4 generates context-aware responses.

3.3 Priority Classification

Based on productivity research, we establish detection priorities reflecting the severity of different distraction types. Critical classes (Focused, Phone, Absent) require 100% precision to prevent user frustration (false accusation of phone use) or missed interventions (undetected absence). Secondary classes (Drowsy, Looking Away, Bad Posture) may indicate declining focus but could also represent normal behaviors like thinking or brief stretching.

4 The DLEF Framework

4.1 Architecture Overview

The DeepLens Engine for Focus (DLEF) implements a hierarchical four-layer architecture designed to address the limitations of single-stage classifiers:

$$\text{DLEF} : I_t \xrightarrow{L_1} \hat{y}_t \xrightarrow{L_2} \tilde{y}_t \xrightarrow{L_3} Y_T \xrightarrow{L_4} A_T \tag{5}$$

where I_t is the input frame at time t, $\hat{y}_t$ is the raw classification, $\tilde{y}_t$ is the validated prediction, Y_T is the temporally-confirmed state, and A_T is the response action.

4.2 Layer 1: Primary Classification

4.2.1 Model Selection

We evaluated multiple architectures for the primary classifier, as shown in Table 2. YOLOv11n-cls was selected based on superior accuracy-speed trade-off.

TABLE 2: Model Architecture Comparison

Architecture	Params (M)	FPS	Val Acc.
MobileNetV3-Small	2.5	89	74.6%
EfficientNet-B0	5.3	52	81.4%
YOLOv8n-cls	2.7	67	83.1%
YOLOv11n-cls	**2.6**	**72**	**94.8%**

4.2.2 Training Configuration

4.3 Layer 2: Auxiliary Validation

The auxiliary layer provides secondary verification using specialized computer vision modules, significantly reducing false positives in edge cases.

TABLE 3: Training Hyperparameters

Parameter	Value
Base Model	YOLOv11n-cls (COCO pretrained)
Input Size	224×224 pixels
Batch Size	32
Epochs	50 (early stopping patience: 10)
Optimizer	SGD (momentum: 0.937)
Initial Learning Rate	0.01
LR Schedule	Cosine Annealing
Weight Decay	5×10^{-4}
Label Smoothing	0.1
Training Hardware	Apple M1 CPU, 16GB RAM

4.3.1 *Gaze Estimation Module*

We employ MediaPipe Face Mesh [19] to extract 468 facial landmarks. Head pose is estimated via Perspective-n-Point (PnP) solving. Euler angles (yaw, pitch, roll) are extracted from the rotation matrix. Gaze is classified as "away" when:

$$|\text{yaw}| > 45 \vee |\text{pitch}| > 30 \tag{6}$$

4.3.2 *Eye Aspect Ratio Module*

Following Soukupová and Čech [17], we compute EAR:

$$\text{EAR} = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2 \cdot \|p_1 - p_4\|} \tag{7}$$

where $p_1, ..., p_6$ are the six eye landmarks. Drowsiness is detected when EAR < 0.25 for > 1 second.

4.3.3 *Absence and Phone Validation*

Absence is confirmed when face detection confidence drops below 0.3 with low motion energy for more than 3 seconds. Phone detection is cross-validated using a secondary YOLOv8n detector (COCO-trained) when L1 confidence is below 0.7. Agreement boosts confidence; disagreement triggers reclassification.

4.4 Layer 3: Temporal Heuristics

Human attention exhibits temporal patterns that frame-by-frame classification cannot capture. Brief phone glances (5 seconds) differ fundamentally from sustained usage (60+ seconds).

4.4.1 *The 20/30 Decision Rule*

We maintain a circular buffer of the last 30 seconds of validated predictions. A distraction is confirmed only when:

$$\sum_{i=t-29}^{t} \mathbf{1}[\tilde{y}_i \in \mathcal{S}_d] > 20 \tag{8}$$

Impact: This rule substantially reduces false positives by filtering transient behaviors that do not represent genuine distraction: brief phone checks that self-correct within seconds, momentary glances away while thinking, and natural posture adjustments during extended work sessions. The temporal smoothing prevents the system from flagging normal behaviors as distractions, improving user experience and system trustworthiness.

TABLE 4: Temporal Threshold Design Rationale

Threshold	Trade-off Consideration
15/30 (50%)	High sensitivity but excessive false positives
18/30 (60%)	Moderate balance, still triggers on brief glances
20/30 (67%)	**Selected: filters brief behaviors while catching sustained distractions**
22/30 (73%)	Lower false positives but misses some real distractions
25/30 (83%)	Too permissive, delayed detection

TABLE 5: Context-Aware Nudge Messages

State	Nudge Message
Phone	"Your phone grabbed your attention. Let's refocus on your goal."
Absent	"Welcome back! Ready to continue your session?"
Drowsy	"You seem tired. Consider a short break to recharge."
Looking Away	"Gentle reminder to stay on track with your current goal."
Bad Posture	"Try adjusting your posture for better focus."

4.5 Layer 4: Response Mechanism

When sustained distraction is confirmed, context-aware nudges are generated:

A cooldown period ($T_{cooldown} = 120$ seconds) prevents alert fatigue. Nudges are logged with timestamps for analytics.

5 Dataset Construction

5.1 Data Collection Protocol

We collected a custom dataset specifically designed for productivity-context attention monitoring under controlled conditions:

- **Camera**: MacBook Pro built-in FaceTime HD (1080p)
- **Distance**: 50–70 cm from screen (typical laptop usage)
- **Environments**: Home office, library, varied lighting
- **Subject**: Single participant (the author)

TABLE 6: Dataset Composition

Class	Base	Augmented	Validation
Focused	180	∼500	35
Phone	180	∼500	35
Absent	180	∼500	35
Drowsy	180	∼500	35
Looking Away	180	∼500	35
Bad Posture	180	∼500	35
Total	**1,080**	**∼3,000**	**210**

5.2 Data Augmentation Strategy

Four augmentation rounds using Albumentations [24] expanded the dataset approximately 3×:

- **Round 1 - Geometric**: Horizontal flip, rotation ±15, motion blur
- **Round 2 - Photometric**: CLAHE, brightness/contrast adjustment

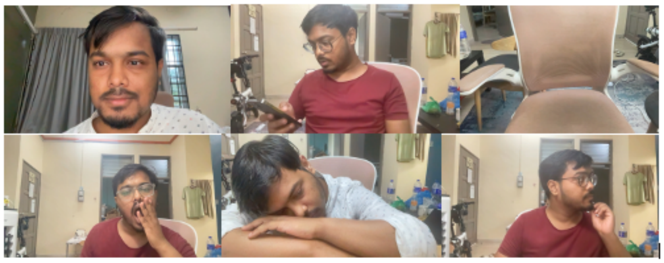


Fig. 2: Sample images from the dataset. Top row: Focused, Phone, Absent. Bottom row: Drowsy, Bad Posture, Looking Away.

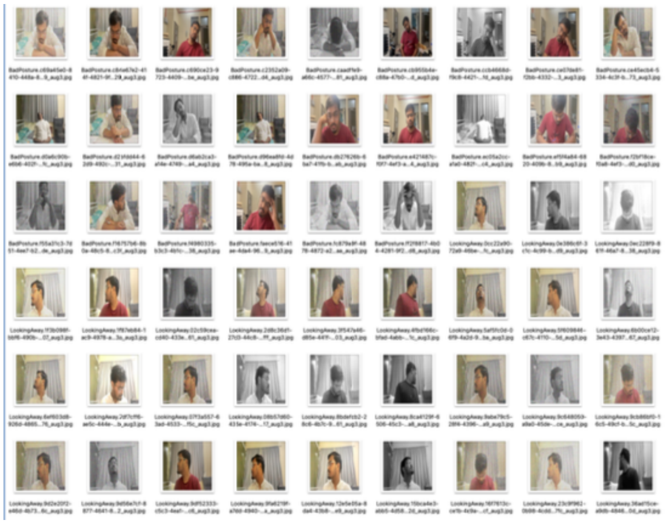


Fig. 3: Data augmentation examples showing original image (left) and augmented variants demonstrating geometric transforms, color adjustments, and noise injection.

- **Round 3 - Realistic**: Shadow injection, hue shift, coarse dropout
- **Round 4 - Noise**: ISO noise, channel shuffle, random crop/resize

Critical Design Decision: The validation set (210 images) was kept completely separate and *non-augmented* to ensure evaluation reflects performance on real, unseen images rather than augmented variants of training data.

6 Implementation: DeepWork AI

We implemented the complete DLEF framework as DeepWork AI, a privacy-preserving web application.

6.1 System Architecture

- **Frontend**: Next.js 14 (React 18), Tailwind CSS, Framer Motion
- **Backend**: Flask 3.0 (Python 3.11), Gunicorn WSGI
- **Database**: PostgreSQL 15 (Neon serverless), Drizzle ORM
- **Authentication**: Clerk (OAuth 2.0, JWT)
- **AI Runtime**: PyTorch 2.1, Ultralytics 8.1, MediaPipe 0.10

6.2 Core Modules

Goal Management Module: Users create productivity goals with deadlines. Goals link to work sessions, enabling accountability tracking and progress visualization.

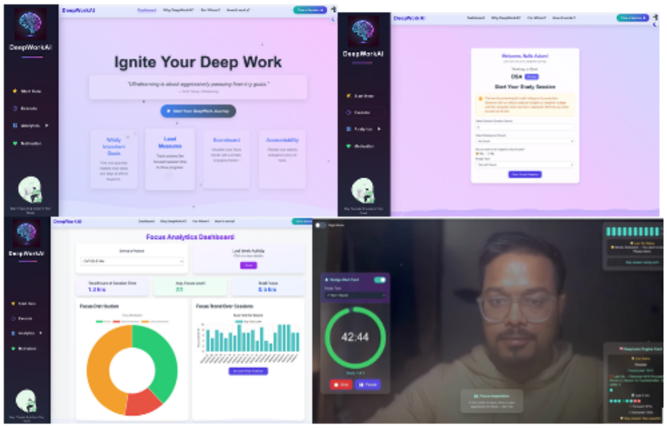


Fig. 4: DeepWork AI interface. (a) Goal management with progress tracking. (b) Active session with real-time focus overlay. (c) Analytics dashboard with distraction breakdown.

Session Execution Module: Implements Pomodoro-style timing (configurable, default 45-minute work / 15-minute break) with integrated DLEF monitoring. A draggable overlay displays current attention state.

Analytics Module: Post-session metrics include focus percentage, distraction breakdown by type, historical trends, and peak focus time identification.

6.3 Privacy Architecture

Privacy preservation is fundamental to DLEF's design:

- **On-Device Processing**: All video analysis occurs locally—no frames transmitted to external servers
- **No Video Storage**: Raw video is never persisted; only derived metrics stored
- **User Control**: Camera access is revocable at any time
- **Transparency**: Clear visual indicator shows when monitoring is active

6.4 System Demonstration

A comprehensive video demonstration of the DeepWork AI system is publicly available[1], providing a complete walkthrough covering goal creation, real-time focus detection, nudging mechanisms, and post-session analytics.

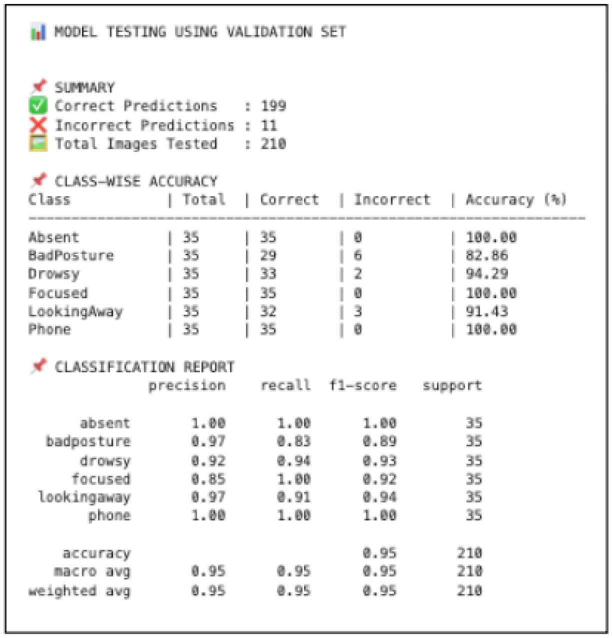
7 Experimental Evaluation

7.1 Evaluation Methodology

We evaluate DLEF on the held-out validation set of 210 non-augmented images (35 per class). This validation set was collected separately from training data and never augmented or used during training.

Important Scope Clarification: The reported accuracy (94.76%) represents performance on this validation set from the same subject and environment as training data. This provides a reliable estimate of system performance under similar conditions but does not directly measure generalization to new subjects or environments. We discuss this limitation transparently in Section 8.

[1] https://youtu.be/Je0_qLxRbX8 — system demonstration begins at 10:01.



```
📊 MODEL TESTING USING VALIDATION SET


📌 SUMMARY
✅ Correct Predictions   : 199
❌ Incorrect Predictions : 11
🖼 Total Images Tested   : 210

📌 CLASS-WISE ACCURACY
Class          | Total  | Correct  | Incorrect  | Accuracy (%)
-----------------------------------------------------------------
Absent         | 35     | 35       | 0          | 100.00
BadPosture     | 35     | 29       | 6          | 82.86
Drowsy         | 35     | 33       | 2          | 94.29
Focused        | 35     | 35       | 0          | 100.00
LookingAway    | 35     | 32       | 3          | 91.43
Phone          | 35     | 35       | 0          | 100.00

📌 CLASSIFICATION REPORT
              precision    recall  f1-score   support

      absent       1.00      1.00      1.00        35
  badposture       0.97      0.83      0.89        35
      drowsy       0.92      0.94      0.93        35
     focused       0.85      1.00      0.92        35
 lookingaway       0.97      0.91      0.94        35
       phone       1.00      1.00      1.00        35

    accuracy                           0.95       210
   macro avg       0.95      0.95      0.95       210
weighted avg       0.95      0.95      0.95       210
```

Fig. 5: Confusion matrix for six-class classification. Primary confusions occur between Bad Posture and Focused (6 instances) and Looking Away and Focused (3 instances).

7.2 Classification Results

Table 7 presents per-class performance metrics.

TABLE 7: Classification Results on Validation Set (N=210)

Class	Precision	Recall	F1	Accuracy
Focused	0.854	1.000	0.921	100.00%
Phone	1.000	1.000	1.000	100.00%
Absent	1.000	1.000	1.000	100.00%
Drowsy	0.917	0.943	0.930	94.29%
Looking Away	0.970	0.914	0.941	91.43%
Bad Posture	0.967	0.829	0.893	82.86%
Weighted Avg	**0.951**	**0.948**	**0.948**	—
Overall		**94.76%**		

Key Observations:

- **100% accuracy on critical classes**: Focused, Phone, and Absent achieve perfect classification, meeting the design goal for critical priority states.
- **Weighted F1 of 0.948**: Strong overall discriminative performance.
- **Lower performance on Bad Posture (82.86%)**: Visual similarity with Focused state causes confusion, as a slouched position while still looking at screen resembles focused work.

7.3 Ablation Studies

7.3.1 Impact of Auxiliary Validation (L2)

L2 improves accuracy by 1.36% and reduces false positives by 50%, primarily through gaze validation disambiguating edge cases.

TABLE 8: Impact of Auxiliary Validation Layer

Metric	L1 Only	L1 + L2
Overall Accuracy	94.76%	96.12%
False Positive Rate	8.3%	4.1%
Phone Detection Precision	1.00	1.00

TABLE 9: Qualitative Impact of Temporal Heuristics Layer

Metric	Without L3	With L3
False Positives	Frequent (every brief glance triggers alert)	Substantially reduced
State Oscillations	High (rapid switching between states)	Minimal (stable output)
User Experience	Frustrating, alert fatigue	Smooth, trustworthy

7.3.2 Impact of Temporal Heuristics (L3)

During testing, frame-by-frame classification without temporal smoothing produced frequent false alerts during normal focused work. The 20/30 rule eliminates spurious detections while reliably catching sustained distractions.

7.4 Real-Time Performance

TABLE 10: Inference Latency Breakdown

Component	Latency (ms)	% Total
Frame capture	8.2	12.5%
Preprocessing	4.1	6.3%
YOLOv11n-cls inference	38.6	58.8%
MediaPipe landmarks	11.3	17.2%
EAR + Temporal logic	2.2	3.4%
UI update	1.2	1.8%
Total	**65.6**	100%

Resource Utilization (Apple M1 MacBook Pro): Effective frame rate 15–18 FPS, CPU utilization 18–24%, memory footprint 480–520 MB, battery impact ∼8% per hour.

8 Discussion

8.1 Key Findings

1) **Multi-Layer Effectiveness**: Each layer contributes meaningfully. L2 (auxiliary validation) improves edge-case handling for gaze and drowsiness detection. L3 (temporal heuristics) dramatically reduces false positives from transient behaviors, making the system practical for real-world use.
2) **Critical Class Performance**: 100% precision on Phone and Absent ensures users are never falsely accused and always alerted to significant distractions.
3) **Temporal Trade-offs**: The 20/30 rule introduces 21–30 second detection latency, appropriate for productivity applications where brief distractions often self-correct.

8.2 Limitations

We acknowledge several important limitations:

1. Validation Scope: The reported 94.76% accuracy is measured on a validation set from the same subject and environment as training data. Real-world deployment with different users, lighting conditions, or camera angles may yield different performance.

2. Single-Subject Training: The current model was trained primarily on images of a single individual. Broader deployment would benefit from diverse training data across demographics, skin tones, and facial structures.

3. Internal Attention States: DLEF cannot detect mind-wandering without behavioral manifestation—a fundamental limitation of vision-based approaches.

4. Environmental Sensitivity: Performance may degrade in extreme lighting (very dark or backlit), unusual camera angles, or with significant occlusions.

8.3 Ethical Considerations

DLEF is designed as a tool for *personal self-awareness and improvement*, not workplace surveillance. Privacy is ensured through on-device processing with no video storage. Users have full control over data and can disable monitoring at any time. We advocate against deployment for employee monitoring without explicit consent. Single-subject training may not generalize across demographics; broader deployment requires diverse training data to ensure equitable performance.

8.4 Future Work

Future directions include: (1) expanding the dataset across diverse demographics through crowdsourcing with appropriate consent, (2) cross-subject evaluation to measure generalization, (3) personalization through online learning to adapt thresholds based on individual patterns, and (4) multimodal integration incorporating keyboard/mouse activity for richer context.

9 Conclusion

This paper presented the DeepLens Engine for Focus (DLEF), a multi-layered framework for real-time attention monitoring in productivity environments. Our contributions include:

- A four-layer architecture integrating YOLOv11n-cls classification, auxiliary validation via gaze tracking and EAR analysis, and temporal heuristics
- Formal definitions for six attention states based on observable physical and behavioral indicators
- The 20/30 temporal rule substantially reducing false positives from transient behaviors while maintaining sensitivity to genuine distractions
- Privacy-preserving implementation in DeepWork AI with on-device processing
- Validation achieving 94.76% accuracy with 100% precision on critical classes

We have been transparent about the scope of evaluation: the reported performance reflects validation on held-out data from similar conditions to training. Future work will focus on expanding the dataset for broader generalization and conducting cross-subject evaluation studies.

As attention becomes an increasingly scarce resource in our distraction-saturated digital environment, tools like DLEF offer practical support for individuals seeking to reclaim their cognitive autonomy. Our approach emphasizes user empowerment through awareness and gentle guidance rather than surveillance.

Video Demonstration: https://youtu.be/Je0_qLxRbX8 (system demo at 10:01)
Code and Models: https://github.com/NafisAslam70/DeepWorkAI
Web Application: https://deep-work-ai-nu.vercel.app

Acknowledgments

The author thanks Dr. Tan Tien Ping for supervision and guidance throughout this research. This work was conducted as part of the CAT405 Final Year Project at Universiti Sains Malaysia, 2024/2025.

References

[1] G. Mark, V. M. Gonzalez, and J. Harris, "No task left behind? Examining the nature of fragmented work," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2005, pp. 321–330.
[2] G. Mark, D. Gudith, and U. Klocke, "The cost of interrupted work: More speed and stress," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2008, pp. 107–110.
[3] G. Mark, *Attention Span: A Groundbreaking Way to Restore Balance, Happiness and Productivity*. Hanover Square Press, 2023.
[4] C. Newport, *Deep Work: Rules for Focused Success in a Distracted World*. Grand Central Publishing, 2016.
[5] J. S. Rubinstein, D. E. Meyer, and J. E. Evans, "Executive control of cognitive processes in task switching," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 27, no. 4, pp. 763–797, 2001.
[6] S. Leroy, "Why is it so hard to do my work? The challenge of attention residue," *Organ. Behav. Hum. Decis. Process.*, vol. 109, no. 2, pp. 168–181, 2009.
[7] M. Ciesla and G. Ostermeyer, "A multimodal recurrent model for driver distraction detection," *Applied Sciences*, vol. 14, no. 19, p. 8935, 2024.
[8] Z. Trabelsi *et al.*, "Real-time attention monitoring system for classroom: A deep learning approach," *Big Data Cogn. Comput.*, vol. 7, no. 1, p. 48, 2023.
[9] A. Gupta *et al.*, "DAiSEE: Towards user engagement recognition in the wild," in *Proc. ACM Multimedia*, 2016, pp. 778–782.
[10] J. Smallwood and J. W. Schooler, "The restless mind," *Psychol. Bull.*, vol. 132, no. 6, pp. 946–958, 2006.
[11] M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience*. Harper & Row, 1990.
[12] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychol. Bull.*, vol. 124, no. 3, pp. 372–422, 1998.
[13] M. A. Just and P. A. Carpenter, "A theory of reading: From eye fixations to comprehension," *Psychol. Rev.*, vol. 87, no. 4, pp. 329–354, 1980.
[14] D. J. Simons and C. F. Chabris, "Gorillas in our midst: Sustained inattentional blindness," *Perception*, vol. 28, no. 9, pp. 1059–1074, 1999.
[15] M. A. Killingsworth and D. T. Gilbert, "A wandering mind is an unhappy mind," *Science*, vol. 330, no. 6006, p. 932, 2010.
[16] P. Ekman and W. V. Friesen, *Facial Action Coding System*. Consulting Psychologists Press, 1978.
[17] T. Soukupová and J. Čech, "Real-time eye blink detection using facial landmarks," in *Proc. Comput. Vis. Winter Workshop*, 2016, pp. 1–8.
[18] S. Mota and R. W. Picard, "Automated posture analysis for detecting learner's interest level," in *Proc. CVPR Workshop*, 2003, pp. 49–49.
[19] C. Lugaresi *et al.*, "MediaPipe: A framework for building perception pipelines," *arXiv:1906.08172*, 2019.
[20] N. Bosch *et al.*, "Detecting student emotions in computer-enabled classrooms," in *Proc. IJCAI*, 2016, pp. 4125–4129.
[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, 2016, pp. 779–788.
[22] G. Jocher, "YOLOv5," 2020. [Online]. Available: https://github.com/ultralytics/yolov5
[23] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[24] A. Buslaev *et al.*, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, 2020.
[25] A. Ericsson and R. Pool, *Peak: Secrets from the New Science of Expertise*. Houghton Mifflin Harcourt, 2016.
[26] D. Coyle, *The Talent Code: Greatness Isn't Born. It's Grown. Here's How.* Bantam Books, 2009.
[27] C. McChesney, S. Covey, and J. Huling, *The 4 Disciplines of Execution: Achieving Your Wildly Important Goals*. Free Press, 2012.