CLEG: Classifying Listener Expertise Through Gaze Patterns During Classical Music Perception

Ko Watanabe*
German Research Center for Artificial Intelligence
Kaiserslautern, Germany
ko.watanabe@dfki.de

Yuki Matsuda*
Okayama University
Okayama, Japan
yukimat@okayama-u.ac.jp

Rinko Hayakawa
Okayama University
Okayama, Japan
rinko@okayama-u.ac.jp

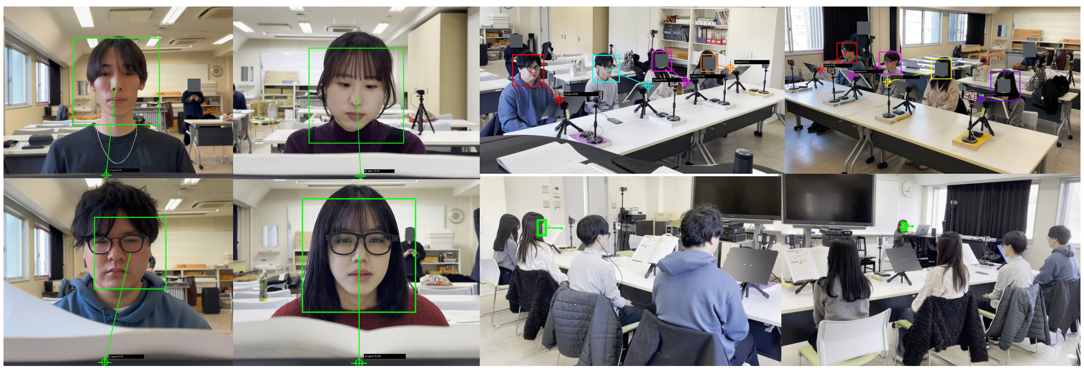


Figure 1: Overview of CLEG: passive sensing of gaze activity during classical music listening, followed by machine-learning-based estimation of listener-side musical states and explainable feedback on the cues driving each prediction.

Abstract

Music education research has largely assessed performers, while listener expertise remains difficult to evaluate objectively without intrusive questionnaires or laboratory-grade sensing. We present CLEG, a system that estimates classical-music listening expertise from RGB-camera-based gaze behavior and explains the cues behind each prediction. From 22 listeners (13 experienced and 9 novice), sliding-window gaze features enabled a Random Forest classifier to separate experienced from novice listeners with an F1 Score of 0.670 at the window level and 0.747 with per-participant majority voting under leave-one-participant-out cross-validation (LOPOCV). SHAP analysis identified maximum fixation duration as the leading cue, suggesting that expertise leaves a recoverable and interpretable signal in commodity-camera gaze data. These results point toward scalable listener-side assessment that can complement performer-focused evaluation in music education.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing**; *Smartphones*; • **Applied computing** → **Education**.

Keywords

Eye Tracking, music, expert sensing, ubiquitous computing

ACM Reference Format:

Ko Watanabe, Yuki Matsuda, and Rinko Hayakawa. 2026. CLEG: Classifying Listener Expertise Through Gaze Patterns During Classical Music Perception. In *Companion of the 2026 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '26), October 11–15, 2026, Shanghai, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3798063.3837164

*These authors contributed equally to this research.

UbiComp Companion '26, Shanghai, China.

ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3798063.3837164

1 Introduction

Classical music education has long emphasized the evaluation of *performers*, and computational systems can now assess pitch, timing, articulation, and expressiveness from performed sound [7, 8, 16, 18]. By contrast, the *listener* side remains difficult to assess objectively, because listening expertise is usually measured through questionnaires, tests, or post-hoc impressions. This limits instructors' ability to observe how musical understanding unfolds during listening and to adapt instruction to each learner.

We address this gap from a ubiquitous-computing perspective by estimating listener expertise from non-contact visual sensing. Prior work suggests that gaze, facial expression, pupil response, and head motion reflect musical attention and expectation [3, 12, 13], while musical expertise shapes visual behavior during music perception [2, 4, 9]. Experienced listeners may anticipate phrase boundaries, harmonic resolutions, and changes in tempo or dynamics, leading to gaze patterns that differ from those of novices. We therefore investigate two research questions:

RQ1 Can classical-music listening expertise be estimated from RGB-camera-based gaze estimation?

RQ2 Which gaze patterns contribute most to estimating novice and experienced classical-music listeners?

To answer these questions, we propose CLEG, a system that records listeners with frontal and in-room cameras and classifies each listener as novice or experienced. We compare a gaze-feature-based binary classification model, and use explainable-AI methods such as SHAP [14] to inspect the cues behind each prediction.

2 Related Work

2.1 Objective Assessment in Music Education

Prior work in music information retrieval and music education has made substantial progress in automatically assessing musical performance from observable properties of the performed sound, such as pitch accuracy, timing deviation, vibrato, articulation, dynamics, and score performance alignment [7, 8, 16, 18]. These studies demonstrate that musical skill can be operationalized through objective features and machine-learning models, but their target is primarily the *performer*: the systems evaluate how accurately or expressively a student sings or plays. In contrast, comparatively little work has addressed the *listener* side of music education, where expertise, preference, comprehension, and engagement are usually measured through questionnaires, tests, or post-hoc ratings rather than continuous and passive sensing.

2.2 Physiological and Behavioral Responses to Music Listening

Music cognition research suggests that listening states are reflected in measurable physiological and behavioral responses. Pupillometry studies have shown that pupil dilation is associated with music-induced arousal, chills, and surprising moments in music [3, 12, 13], while eye-tracking studies of score reading indicate that musical expertise affects fixation patterns and visual auditory integration during music reading [2, 4]. Neuroscientific studies using EEG and fMRI further show that musicians and non-musicians process harmonic structure and passive music listening differently [9, 17]. However, these findings are typically obtained with laboratory-grade eye trackers, EEG systems, or MRI scanners, and they do not directly provide a scalable method for estimating listener preference or expertise in everyday music-listening contexts.

2.3 Wearable Sensing, Non-contact Sensing, and Explainable User-State Estimation

Affective-computing research has used physiological and behavioral signals to infer arousal, valence, stress, engagement, and multimedia preferences [10, 15, 22, 24, 25, 27]. Recent work has explored portable EEG for music preference prediction [5], while camera-based gaze and facial analysis enable attention and affect estimation with commodity sensors [11, 23, 28]. Explainable AI methods such as LIME, SHAP, and attention rollout further help audit whether predictions rely on meaningful cues rather than spurious correlations [1, 14, 19]. However, existing work remains fragmented across performer assessment, laboratory music cognition, and generic affect estimation, leaving interpretable, non-contact estimation of classical-music listener expertise underexplored.

3 Data Collection

We collected a non-contact, multi-camera dataset of listeners during classical music perception in a setting that approximates an ordinary classroom or music-appreciation course. This section describes the participants, the recording apparatus, the musical stimuli, the experimental procedure, and the labeling protocol used to derive listener-expertise annotations.

3.1 Participants

We recruited 22 university students (11 male and 11 female), aged between 21 and 24 years (mean = 22.5). Based on a pre-experiment questionnaire (subsection 3.3) and an assessment by a music-education instructor, participants were assigned to two expertise groups: an *experienced* group of 13 students who were music majors or had instrumental training, and a *novice* group of 9 students with no formal musical training. All participants provided written informed consent prior to the session.

3.2 Experimental Conditions

As illustrated in Figure 1, each participant was seated on a chair and recorded simultaneously by a frontal face camera and an in-room fixed camera. The frontal camera was an iPhone positioned directly in front of the participant, capturing the face for fine-grained gaze and facial analysis. The in-room cameras were fixed cameras placed at the four corners of the room, capturing the participant's whole body and the surrounding environment and providing the wider field of view needed. All sessions used the same loudspeaker, and the music was reproduced under identical acoustic conditions (volume, speaker position, and room) across participants in order to avoid confounds arising from differences in playback.

We prepared a set of Western classical pieces as listening stimuli. In this work we focus on the *Farandole* from Georges Bizet's *L'Arlésienne* Suite No. 2. We trimmed the music into three-minute segments, and participants listened to the same piece twice. The pieces included salient, anticipatable events, such as phrase boundaries, harmonic resolutions, and changes in tempo or dynamics, where experienced listeners may differ most from novices.

3.3 Experimental Procedure

Sessions were conducted in a lecture room. After receiving an explanation of the study and providing written informed consent, participants completed a pre-experiment questionnaire covering age, gender, nationality, and musical experience. They were then seated and listened to the stimuli while looking either at a printed score or straight ahead. To keep behavior as natural as possible and to avoid task-induced changes in gaze, the instructions were kept minimal: participants were simply asked to listen naturally and were given no explicit analysis or rating task during listening. After each piece, participants completed a short impression questionnaire to obtain subjective ratings of their listening experience.

3.4 Annotation and Labeling

Listener-expertise labels were derived from the pre-experiment questionnaire (musical history, academic major, and self-reported sight-reading ability) together with an evaluation by a music-education instructor, yielding the two expertise groups described in subsection 3.1. These group assignments provide the ground-truth labels for the binary novice/experienced classification models used in our experiments (RQ1), as described in section 4. For model training, the recordings were segmented into fixed-length windows using a sliding-window scheme, so that each window inherits the expertise label of its participant and serves as an independent sample.

4 Methodology

CLEG estimates a listener's musical expertise from gaze behavior recorded during classical-music perception. The pipeline consists of four stages: (i) temporal preprocessing and gaze estimation, (ii) gaze-event extraction, (iii) sliding-window feature extraction, and (iv) expertise classification with eight models. We evaluate the binary classification of novice and experienced listeners under a participant-independent protocol. Finally, we attach an explainability method to each model and evaluate under a participant-independent protocol.

4.1 Preprocessing and Gaze Estimation

We first aligned all videos to a common timeline by manually trimming one reference clip per recording group and synchronizing the remaining videos to it using audio. We then retained only the classical-music playback segments. Gaze was estimated from the trimmed videos using Gaze–LLE [20], which outputs per-frame records (t, x, y, c, o): timestamp, normalized gaze coordinates, confidence, and in-frame indicator. These per-participant streams form the input to all subsequent feature-extraction steps.

4.2 Gaze-Event Extraction

From each gaze stream we extracted area-of-exploration (AOE) movements, fixations, and saccades.

AOE characterizes visual exploration through the magnitude and direction of gaze movement between consecutive samples. For each pair of successive gaze points, we computed the inter-sample displacement (point-to-point amplitude) and its movement direction, following standard sample-level eye-movement measures [6]. These quantities summarize the spatial dynamics of gaze over time.

Fixations were detected using the dispersion-threshold (I-DT) algorithm [21], in which consecutive gaze points are grouped into a fixation as long as their spatial dispersion stays below a threshold for at least a minimum duration. Following this definition, we used a dispersion threshold of 0.08 in normalized coordinates and a minimum duration of 0.10 s, and recorded each fixation's timing, duration, mean position, dispersion, and sample count.

Saccades were defined as transitions between successive fixations [21]. For each saccade, we measured its amplitude as the spatial offset between the two fixation centers and its velocity as the amplitude divided by the transition duration, following standard eye-movement measures [6]. These quantities capture shifts of attention between regions.

4.3 Sliding-Window Feature Extraction

Each gaze stream was divided into overlapping 5.0 s windows with a 2.5 s step (50% overlap). For each window, we extracted features from x, y, AOE amplitude a, AOE angle θ, fixations, and saccades, following previous work [26]. These included time-domain statistics, FFT features over 0.1 to 1.0 Hz, cross-dimensional descriptors, and quality measures such as gaze-point count and effective sampling rate. Each window yielded one feature vector and inherited its participant's expertise label.

4.4 Expertise-Classification Models

The classification target is the listener's musical expertise. Ground-truth labels were derived as described in subsection 3.4 from the pre-experiment questionnaire (years of instrumental training, academic major, and self-reported sight-reading ability) together with an assessment by a music-education instructor, yielding the two expertise groups defined in subsection 3.1 (Experienced and Novice). For our experiments, we use novice and experienced listeners as a binary classification target. In the feature-based pipeline, the per-window feature vectors of subsection 4.3 are fed to classical machine-learning classifiers. This pipeline is interpretable at the feature level and provides the basis for the SHAP analysis.

4.5 Explainability

A central goal of CLEG is to ensure that predictions rest on musically and physiologically plausible evidence rather than spurious correlations. For the feature-based classifier we compute SHAP values [14], which quantify the contribution of every gaze feature to each prediction. Because samples are time-localized windows, the attributions can be aggregated per participant and per musical moment, revealing which features most strongly distinguish experienced from novice listeners, and at which points in the piece.

4.6 Evaluation Protocol

We evaluate with participant-independent leave-one-participant-out-cross-validation (LOPOCV): in each fold all windows of one participant form the test set and the remaining participants form the training set, so that no participant appears in both. This prevents identity leakage across overlapping windows and measures generalization to unseen listeners. We report classification performance for novice and experienced listeners. Beyond predictive performance, we examine the features identified by SHAP and the segments in which experienced and novice listeners differ most, linking model behavior back to interpretable gaze patterns.

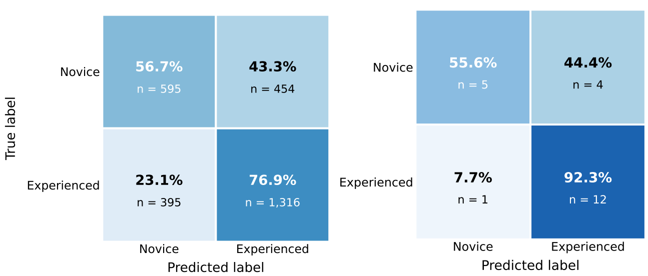
5 Results and Discussion

Table 1 reports the classification performance under the participant-independent LOPOCV protocol. Using sliding-window features, the Random Forest classifier achieved an F1 Score of 0.670 at the window level (WL), increasing to 0.747 with majority voting (MV). Other feature-based models showed similar MV performance, suggesting that participant-level aggregation reduces window-level noise. For deep-learning models, the best performance was obtained by MobileNetV3 with an F1 Score of 0.571 at WL and 0.705 at MV. All results exceed the chance baseline (50%), indicating that listening expertise is recoverable from RGB-camera-based gaze estimation alone (**RQ1**).

Figure 2 shows the confusion matrices for the WL and MV Random Forest classifiers. In both settings the experienced class is recognized with high reliability (76.9% at the window level and

Table 1: Classification performance under participant-independent LOPOCV with undersampling. For each model, both window-level (WL) and participant-level majority voting (MV) scores are reported.

Model	Level	Acc.	Prec.	Rec.	F1 Score
Extra Trees	WL	0.671	0.649	0.646	0.647
	MV	0.773	0.792	0.739	0.747
Linear SVM	WL	0.638	0.615	0.614	0.614
	MV	0.773	0.792	0.739	0.747
Logistic Regression	WL	0.643	0.622	0.622	0.622
	MV	0.773	0.792	0.739	0.747
Random Forest	WL	0.692	0.672	0.668	**0.670**
	MV	0.773	0.792	0.739	**0.747**
LSTM	WL	0.611	0.604	0.610	0.602
	MV	0.545	0.514	0.513	0.509
MobileNet	WL	0.605	0.590	0.593	0.590
	MV	0.682	0.677	0.645	0.646
MobileNetV2	WL	0.572	0.561	0.564	0.559
	MV	0.591	0.571	0.568	0.569
MobileNetV3	WL	0.589	0.570	0.572	0.571
	MV	0.727	0.724	0.701	0.705



(a) Window-level (WL) **(b) Majority vote (MV)**

Figure 2: Confusion matrices for the window-level and majority-voting Random Forest classifiers.

92.3% under MV), whereas the novice class is identified less consistently (56.7% and 55.6%, respectively). This asymmetry indicates that the visual behavior of experienced listeners is more distinctive and easier to separate, which is consistent with the idea that expertise produces structured, anticipatory gaze patterns rather than the more variable behavior of novices.

Figure 3 presents the SHAP results for the Novice class (**RQ2**). Among the top-ranked features, fixation-duration descriptors (maximum fixation duration) emerge as strong contributors to the prediction, together with saccade-duration features. The direction of these attributions shows that experienced and novice listeners occupy clearly different regions of the feature distribution. A plausible interpretation is that novices, who cannot fluently follow the printed score with their eyes, tend to fixate on a single point for longer, yielding larger maximum fixation durations, whereas experienced listeners distribute their gaze and shift it in anticipation of upcoming musical structure. The dominance of musically and

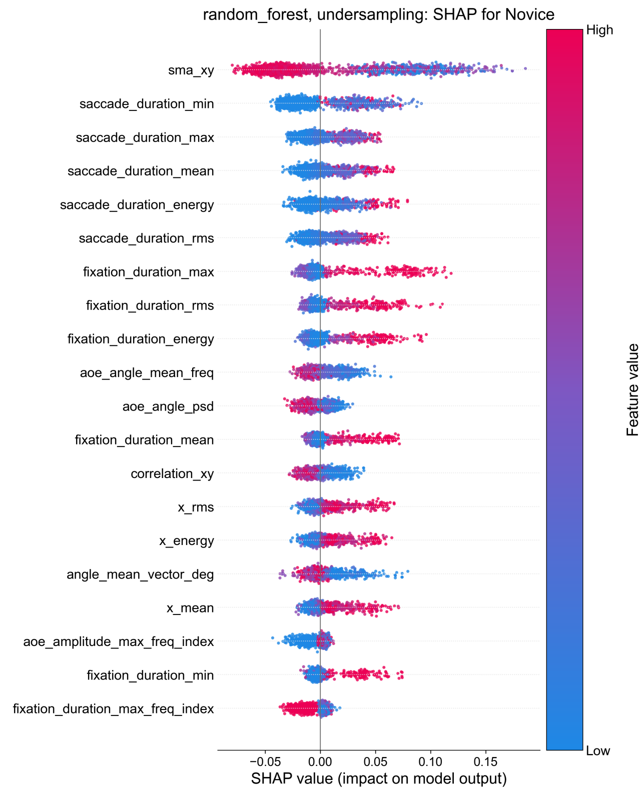


Figure 3: SHAP beeswarm plot for the Novice class.

physiologically plausible gaze features, rather than spurious cues, supports the validity of the model's decisions.

Taken together, these results suggest that classical-music listening expertise can be estimated from non-contact, RGB-camera-based gaze information, and that the cues driving the estimation correspond to interpretable differences in fixation and saccade behavior between experienced and novice listeners in our score-viewing setting.

6 Conclusion

We presented CLEG, a protocol for estimating classical-music listening expertise from non-contact, RGB-camera-based gaze sensing. Using Gaze–LLE gaze streams and sliding-window gaze features, a Random Forest classifier separated experienced from novice listeners with an F1 Score of 0.670 at the window level and 0.747 with per-participant majority voting, exceeding the chance baseline under LOPOCV. SHAP analysis indicated that maximum fixation duration was the main contributor, suggesting a difference in how novices and experienced listeners distribute gaze during listening.

Acknowledgments

This work was supported in part by JSPS KAKENHI Grant Number JP26K21206 and KAKENHI Safety-Net Grant funded by Okayama University. Claude Opus 4.8 was used to assist with manuscript writing; all generated text was reviewed and revised by the authors.

References

[1] Samira Abnar and Willem Zuidema. 2020. Quantifying Attention Flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4190–4197. doi:10.18653/v1/2020.acl-main.385

[2] Véronique Drai-Zerbib, Thierry Baccino, and Emmanuel Bigand. 2012. Sight-Reading Expertise: Cross-Modality Integration Investigated Using Eye Tracking. *Psychology of Music* 40, 2 (2012), 216–235. doi:10.1177/0305735610394710

[3] Bruno Gingras, Manuela M. Marin, Eva Puig-Waldmüller, and W. Tecumseh Fitch. 2015. The Eye Is Listening: Music-Induced Arousal and Individual Differences Predict Pupillary Responses. *Frontiers in Human Neuroscience* 9 (2015), 619. doi:10.3389/fnhum.2015.00619

[4] Thomas W. Goolsby. 1994. Eye Movement in Music Reading: Effects of Reading Ability, Notational Complexity, and Encounters. *Music Perception* 12, 1 (1994), 77–96. doi:10.2307/40285756

[5] Zhiyu He, Jiayu Li, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2024. Introducing EEG Analyses to Help Personal Music Preference Prediction. arXiv:2404.15753 [cs.IR]

[6] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures.* Oxford University Press.

[7] Jiawen Huang, Yun-Ning Hung, Ashis Pati, Siddharth Kumar Gururani, and Alexander Lerch. 2020. Score-Informed Networks for Music Performance Assessment. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*. 908–915.

[8] Norihiro Kato, Eita Nakamura, Kyoko Mine, Orie Doeda, and Masanao Yamada. 2023. Computational Analysis of Audio Recordings of Piano Performance for Automatic Evaluation. In *Responsive and Sustainable Educational Futures (Lecture Notes in Computer Science)*. Springer, 586–592. doi:10.1007/978-3-031-42682-7_46

[9] Stefan Koelsch, Björn-Helmer Schmidt, and Julia Kansok. 2002. Effects of Musical Expertise on the Early Right Anterior Negativity: An Event-Related Brain Potential Study. *Psychophysiology* 39, 5 (2002), 657–663. doi:10.1017/S0048577202010508

[10] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A Database for Emotion Analysis Using Physiological Signals. *IEEE Transactions on Affective Computing* 3, 1 (2012), 18–31. doi:10.1109/T-AFFC.2011.15

[11] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye Tracking for Everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2176–2184. doi:10.1109/CVPR.2016.239

[12] Bruno Laeng, Lise Mette Eidet, Unni Sulutvedt, and Jaak Panksepp. 2016. Music Chills: The Eye Pupil as a Mirror to Music's Soul. *Consciousness and Cognition* 44 (2016), 161–178. doi:10.1016/j.concog.2016.07.009

[13] Hsin-I Liao, Makoto Yoneya, Makio Kashino, and Shigeto Furukawa. 2018. Pupillary Dilation Response Reflects Surprising Moments in Music. *Journal of Eye Movement Research* 11, 2 (2018), 13. doi:10.16910/jemr.11.2.13

[14] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

[15] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing* 10, 1 (2019), 18–31. doi:10.1109/TAFFC.2017.2740923

[16] Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga. 2006. An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitch Interval Accuracy and Vibrato Features. In *Interspeech 2006*. doi:10.21437/Interspeech.2006-474

[17] Takashi Ohnishi, Hiroshi Matsuda, Takashi Asada, Makoto Aruga, Makiko Hirakata, Masami Nishikawa, Asako Katoh, and Etsuko Imabayashi. 2001. Functional Anatomy of Musical Perception in Musicians. *Cerebral Cortex* 11, 8 (2001), 754–760. doi:10.1093/cercor/11.8.754

[18] Kumar Ashis Pati, Siddharth Gururani, and Alexander Lerch. 2018. Assessment of Student Music Performances Using Deep Neural Networks. *Applied Sciences* 8, 4 (2018), 507. doi:10.3390/app8040507

[19] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144. doi:10.1145/2939672.2939778

[20] Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M. Rehg. 2025. Gaze-LLE: Gaze Target Estimation via Large-Scale Learned Encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 28874–28884.

[21] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Palm Beach Gardens, Florida, USA) *(ETRA '00)*. Association for Computing Machinery, New York, NY, USA, 71–78. doi:10.1145/355017.355028

[22] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 400–408. doi:10.1145/3242969.3242985

[23] Vijul Shah, Ko Watanabe, Brian Moser, and Andreas Dengel. 2025. PupilSense: A Novel Application for Webcam-Based Pupil Diameter Estimation. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications (ETRA '25)*. Association for Computing Machinery, New York, NY, USA, Article 11, 8 pages. doi:10.1145/3715669.3723125

[24] Noriyuki Tanaka, Ko Watanabe, Shoya Ishimaru, Andreas Dengel, Shingo Ata, and Manato Fujimoto. 2024. Concentration Estimation in Online Video Lecture Using Multimodal Sensors. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Melbourne VIC, Australia) *(UbiComp '24)*. Association for Computing Machinery, New York, NY, USA, 71–75. doi:10.1145/3675094.3677587

[25] Ko Watanabe, Tanuja Sathyanarayana, Andreas Dengel, and Shoya Ishimaru. 2023. EnGauge: Engagement Gauge of Meeting Participants Estimated by Facial Expression and Deep Neural Network. *IEEE Access* 11 (2023), 52886–52898. doi:10.1109/ACCESS.2023.3279428

[26] Ko Watanabe, Yusuke Soneda, Yuki Matsuda, Yugo Nakamura, Yutaka Arakawa, Andreas Dengel, and Shoya Ishimaru. 2021. DisCaaS: Micro Behavior Analysis on Discussion by Camera as a Sensor. *Sensors* 21, 17 (2021). doi:10.3390/s21175719

[27] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. 2014. The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98. doi:10.1109/TAFFC.2014.2316163

[28] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-Based Gaze Estimation in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4511–4520. doi:10.1109/CVPR.2015.7299081