Label-Efficient Multimodal Microsleep Forecasting via Interpretable Disagreement-Driven Active Learning

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Forecasting microsleep events from multimodal EEG/EOG signals is important for early vigilance assessment, but expert annotation is costly. This work investigates whether active learning (AL) can reduce labeling effort while preserving forecasting performance. Using the Maintenance of Wakefulness Test (MWT) dataset, we evaluate 1 s and 2 s ahead microsleep forecasting under strict Leave-One-Subject-Out validation. We benchmark four pool-based AL strategies and propose **Multimodal Disagreement Active Learning (MD-AL)**, an interpretable acquisition strategy combining EEG–EOG disagreement, decision-boundary uncertainty, and rare-event prioritization.
Results show that AL recovers approximately 95% of fully supervised performance using only 1–2% labeled samples. MD-AL performs competitively with established AL methods while providing a physiology-aware and explainable acquisition mechanism. Ablation analysis shows that combining multimodal disagreement with uncertainty and rare-event prioritization yields a more stable query signal than disagreement alone. These findings support active learning as a practical route toward label-efficient and interpretable microsleep forecasting.

Keywords: Microsleep Forecasting · Active Learning · Multimodal Fusion

1 Introduction

Microsleep events are brief, involuntary lapses of attention that can compromise safety in driving, transportation, and other sustained-attention settings [1–3]. EEG-based approaches have been widely studied for microsleep and vigilance-state detection using spectral, feature-based, and deep learning methods [4–7, 12]. However, forecasting microsleep before onset remains more challenging, with performance typically degrading as the prediction horizon increases [9, 10, 13, 11]. Reliable early-warning systems therefore require models that identify subtle pre-event physiological changes while remaining robust to subject variability.

A major barrier is the cost of expert annotation. Microsleep labels require careful review of EEG/EOG recordings and behavioral cues, making large-scale supervised training expensive. In contrast, unlabeled physiological recordings are

often more readily available. This motivates label-efficient learning approaches that preserve performance while reducing the number of samples requiring expert annotation.

Active learning (AL) addresses this problem by selectively querying informative unlabeled samples. Prior work in EEG classification, brain–computer interfaces, and driver monitoring has shown that uncertainty, diversity, and representativeness-based sampling can reduce labeling requirements [14–16]. Batch-mode AL further emphasizes the need to avoid redundant or outlier samples when multiple queries are selected simultaneously [17]. However, most existing AL methods are modality-agnostic and are mainly designed for classification, calibration, or state estimation rather than subject-independent microsleep forecasting across future horizons.

In multimodal physiological monitoring, disagreement between EEG and EOG predictions may identify transitional vigilance states, conflicting physiological evidence, or difficult decision regions where annotation is especially valuable. Motivated by this observation, we propose **Multimodal Disagreement Active Learning (MD-AL)**, an interpretable acquisition strategy that combines EEG–EOG disagreement, decision-boundary uncertainty, and rare-event prioritization. Each component has a direct physiological or statistical interpretation, making the query process explainable by design.

Using the publicly available MWT dataset, this work evaluates whether AL can reduce annotation requirements for 1 s and 2 s ahead microsleep forecasting under strict Leave-One-Subject-Out validation. The main contributions are: (i) an interpretable multimodal disagreement-driven AL strategy for EEG/EOG-based microsleep forecasting; (ii) evidence that competitive subject-independent forecasting can be achieved with substantially fewer labels than full supervision; and (iii) benchmarking, statistical validation, and ablation analysis against four standard pool-based AL strategies.

2 Methodology

2.1 Dataset and Experimental Setup

Experiments used the public Maintenance of Wakefulness Test (MWT) dataset [18], containing EEG/EOG recordings from 76 participants. Signals were sampled at 200 Hz and segmented into overlapping 1 s windows with a 0.5 s stride. Labels for microsleep episode, microsleep candidate, and drowsiness were merged into one positive class, while wakefulness formed the negative class. Forecasting labels were generated by shifting labels by two windows for 1 s ahead prediction and four windows for 2 s ahead prediction. Features were z-score normalized per subject.

Each sample was represented by a 71-dimensional multimodal feature vector, including EEG spectral, Hjorth, entropy, and mutual-information descriptors; EOG temporal, spectral, entropy, and mutual-information descriptors; and 12 cross-modal mutual-information features capturing EEG–EOG coupling [10, 8].

Following prior multimodal microsleep forecasting work, the classifier was fixed as an Early Fusion (MI Cross) LightGBM model [8]. This fixed-classifier design ensures that differences in performance can be attributed to the acquisition strategy rather than changes in model architecture.

A strict Leave-One-Subject-Out (LOSO) protocol was used. In each fold, one subject was held out for testing and the remaining subjects formed the active learning pool.

Data use declaration: The dataset is publicly available through Zenodo under a Creative Commons Attribution 4.0 International license; no additional private data were collected or used in this study.

2.2 Active Learning Framework

Only a small balanced seed set, approximately 1% of the training pool, was initially labeled. Samples were then queried iteratively until reaching the budgets

$$\mathcal{B} = \{1, 2, 5, 10, 20, 40, 60, 80, 100\}\%.$$

For each budget b, the labeled set size was $n_b = \max(20, \lfloor bN_{\text{train}} \rfloor)$, where N_{train} is the number of training samples in the current fold. After each acquisition step, the Early Fusion model was retrained and evaluated on the held-out subject.

Four pool-based baselines were evaluated: Random Sampling, Uncertainty Sampling using predictive entropy $H(p) = -p \log p - (1-p)\log(1-p)$, Diversity Sampling using greedy farthest-first selection, and Query-by-Committee (QBC) using the variance of predictions from three LightGBM models, $D(x) = \text{Var}(p_1(x), p_2(x), p_3(x))$.

2.3 Multimodal Disagreement Active Learning

The proposed MD-AL strategy uses disagreement between EEG and EOG predictions as a domain-aware acquisition signal. For each unlabeled sample x, synchronized EEG-only, EOG-only, and multimodal fusion models produce $p_{\text{EEG}}(x)$, $p_{\text{EOG}}(x)$, and $p_{\text{fusion}}(x)$. The acquisition score is

$$S(x) = |p_{\text{EEG}}(x) - p_{\text{EOG}}(x)| \cdot 4p_{\text{fusion}}(x)(1 - p_{\text{fusion}}(x)) \cdot (0.5 + p_{\text{fusion}}(x)).$$

The first term captures cross-modal disagreement, the second prioritizes samples near the fusion-model decision boundary, and the third introduces a mild rare-event bias toward likely microsleep samples. Thus, MD-AL is interpretable by design: its score reflects modality conflict, uncertainty, and minority-event prioritization. Samples were ranked by $S(x)$, followed by a greedy diversity step among top candidates to reduce batch redundancy. Query scores were computed without access to unlabeled ground-truth labels. When SMOTE was used, it was applied to the full multimodal feature representation before modality-specific partitioning.

2.4 Evaluation Protocol

Ablations evaluated disagreement alone, disagreement plus boundary uncertainty, and the full MD-AL score. All methods used the same Early Fusion LightGBM classifier for final evaluation; EEG-only and EOG-only models were used only for MD-AL acquisition scoring. Class imbalance was handled using class weighting and SMOTE where appropriate.

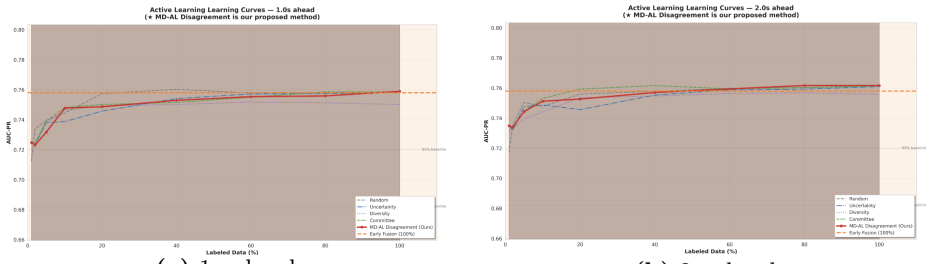
Performance was evaluated using AUC-PR as the primary metric, with AUC-ROC, F1-score, Recall@90% Specificity, and Brier score as secondary metrics. Statistical comparisons against the fully supervised baseline used one-sided Wilcoxon signed-rank tests, while differences among AL strategies were assessed using Friedman tests followed by Holm-corrected pairwise Wilcoxon tests. Label efficiency was summarized using the Area Under the Learning Curve (AULC), with significance set at $p < 0.05$.

3 Results

We compare MD-AL against Random, Uncertainty, Diversity, and QBC under LOSO validation for 1 s and 2 s forecasting. The fully supervised Early Fusion model trained with 100% labeled data serves as the reference baseline.

3.1 Learning Curves

Figure 1 shows AUC-PR learning curves across both horizons. Performance generally improves with labeling budget and converges toward the fully supervised baseline. MD-AL follows a stable trajectory and remains competitive with the strongest baselines. Across both horizons, approximately 95% of fully supervised performance is achieved using only 1–2% labeled samples. Although QBC occasionally reaches the baseline with fewer labels, MD-AL achieves comparable final performance while providing an interpretable multimodal acquisition mechanism.

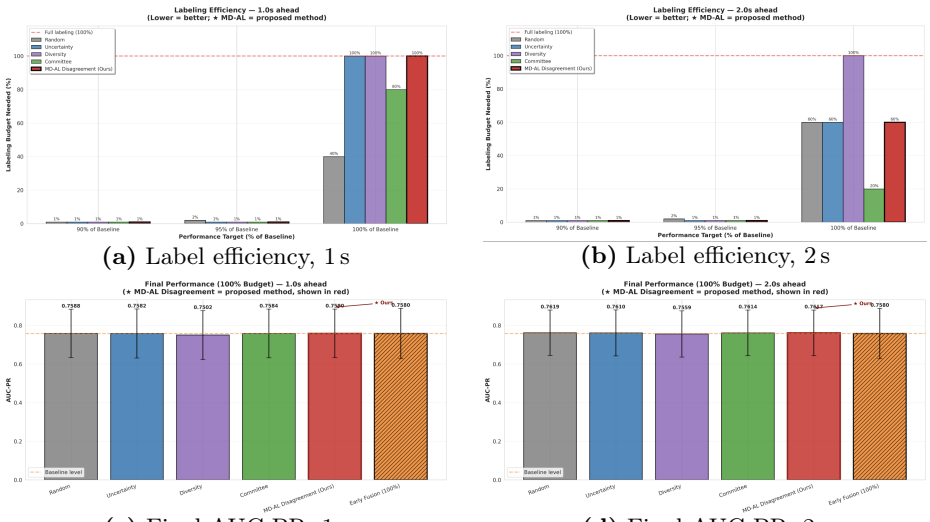


(a) 1 s ahead **(b)** 2 s ahead

Fig. 1. AUC-PR learning curves under LOSO validation. The dashed line denotes the fully supervised Early Fusion baseline.

3.2 Label Efficiency and Final Performance

Figure 2 summarizes annotation efficiency and final AUC-PR. At near-baseline operating points, all AL strategies substantially reduce labeling requirements. MD-AL reaches approximately 95% of the fully supervised baseline using only 1–2% labeled data across both horizons. At the stricter 100% baseline threshold, QBC generally requires the smallest labeling budget, while MD-AL remains comparable to Random and Uncertainty.



(a) Label efficiency, 1 s **(b)** Label efficiency, 2 s

(c) Final AUC-PR, 1 s **(d)** Final AUC-PR, 2 s

Fig. 2. Label efficiency and final performance. Panels (a,b) show the minimum labeling budget required to reach 90%, 95%, and 100% of the fully supervised baseline. Panels (c,d) show final AUC-PR.

At the largest labeling budget, performance differences among methods are small. MD-AL achieves one of the highest final AUC-PR values for both horizons and remains numerically comparable to the fully supervised baseline, indicating that multimodal disagreement preserves predictive performance while improving interpretability.

3.3 Statistical Validation and Ablation

Table 1 summarizes the statistical validation of the benchmark AL strategies at the 20% labeling budget; MD-AL is evaluated separately through the ablation analysis in Fig. 3. Wilcoxon tests show that Random and QBC were not significantly worse than full supervision for the 1 s horizon, whereas Diversity and Uncertainty were significantly lower. Friedman tests further indicate significant differences among AL strategies for both horizons, confirming that acquisition strategy affects label efficiency.

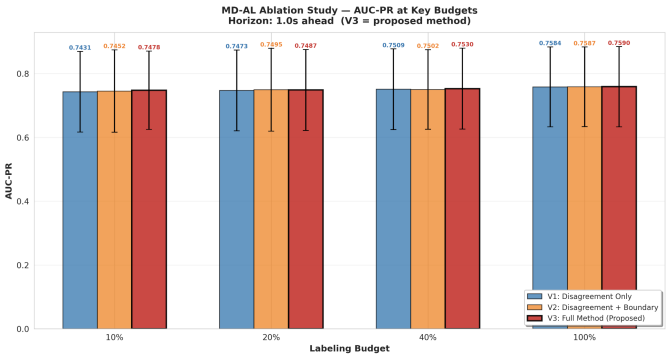


Fig. 3. Ablation study of the MD-AL acquisition score at key labeling budgets for the 1 s prediction horizon.

Table 1. Statistical validation at 20% labeling. Wilcoxon tests compare benchmark AL strategies against full supervision for the 1 s horizon; MD-AL is assessed separately in Fig. 3.

Analysis	Method/Horizon	AUC-PR / χ^2	p
Wilcoxon	Random, 1 s	0.7574	0.4287
Wilcoxon	QBC, 1 s	0.7502	0.1638
Wilcoxon	Diversity, 1 s	0.7493	0.0303
Wilcoxon	Uncertainty, 1 s	0.7458	0.0092
Friedman	1 s horizon	$\chi^2 = 18.09$	4.22×10^{-4}
Friedman	2 s horizon	$\chi^2 = 13.26$	4.10×10^{-3}

Figure 3 shows the MD-AL ablation. Disagreement alone provides an informative signal, but adding boundary-awareness improves performance by prioritizing decision-boundary samples. The full MD-AL formulation, including rare-event prioritization, yields the most consistent performance across budgets, indicating that cross-modal disagreement, uncertainty, and class-aware prioritization contribute complementary information.

4 Discussion

This study demonstrates that active learning can substantially reduce annotation requirements for multimodal microsleep forecasting while preserving subject-independent performance. Across both 1 s and 2 s horizons, near-baseline AUC-PR was achieved with a small fraction of labeled data, indicating that the fully supervised decision boundary can be approximated efficiently under constrained annotation budgets.

A key finding is that simple acquisition strategies are highly competitive. Random and QBC sampling achieved performance statistically comparable to the fully supervised baseline at reduced labeling budgets, suggesting that the multimodal feature representation provides a well-structured decision space. In this setting, the role of active learning is not necessarily to improve final perfor-

mance beyond full supervision, but to reach comparable performance with fewer expert annotations.

MD-AL provides a complementary advantage through interpretability. Unlike generic uncertainty- or committee-based acquisition rules, it explicitly combines EEG–EOG disagreement, decision-boundary proximity, and rare-event prioritization. Cross-modal disagreement highlights samples where neural and ocular signals provide conflicting evidence; the boundary term targets uncertain samples; and the rare-event term mildly emphasizes likely microsleep samples in an imbalanced setting. Thus, MD-AL offers an acquisition mechanism that is explainable by design.

Although MD-AL does not consistently outperform the strongest baseline at every budget, it remains competitive across horizons while producing a transparent query score. This is important for biomedical applications, where annotation requests should be understandable to domain experts. The ablation further supports the proposed design, showing that disagreement alone is useful, but combining it with boundary-awareness and rare-event prioritization yields a more stable acquisition function.

This work has limitations. Experiments were conducted on a single public dataset, and external validation is needed to assess generalization. The acquisition function is manually designed; future work may investigate adaptive weighting of its components and extension to additional physiological modalities or longer prediction horizons.

5 Conclusion

This work establishes active learning as a practical approach for reducing annotation burden in multimodal microsleep forecasting. The proposed MD-AL strategy further shows that EEG–EOG disagreement can serve as an interpretable and physiologically meaningful signal for label-efficient biosignal annotation. While MD-AL does not consistently outperform all baselines, it provides competitive forecasting performance together with an explainable acquisition mechanism tailored to multimodal physiological signals.

References

1. Young, T.B.: Epidemiology of daytime sleepiness: definitions, symptomatology, and prevalence. Journal of Clinical Psychiatry **65**(Suppl. 16), 12–16 (2004)
2. Ford, E.S., Wheaton, A.G., Cunningham, T.J., Giles, W.H., Croft, J.B.: Trends in insomnia and excessive daytime sleepiness among U.S. adults from 2002 to 2012. Sleep Medicine **16**(3), 372–378 (2015). https://doi.org/10.1016/j.sleep.2014.12.008
3. Zhang, Z., Wang, C., Li, Y., Chen, B.: A vehicle active safety model: vehicle speed control based on driver vigilance detection using wearable EEG and sparse representation. Sensors **16**(2), 242 (2016). https://doi.org/10.3390/s16020242
4. Peiris, M.R., Jones, R.D., Davidson, P.R., Bones, P.J.: Detecting behavioral microsleeps from EEG power spectra. In: Proceedings of the Annual International

Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5723–5726 (2006). https://doi.org/10.1109/IEMBS.2006.260411
5. Ayyagari, S.S.D.P., Jones, R.D., Weddell, S.J.: Detection of microsleep states from the EEG: a comparison of feature reduction methods. Medical & Biological Engineering & Computing **59**(7–8), 1643–1657 (2021). https://doi.org/10.1007/s11517-021-02386-y
6. Balaji, A., Tripathi, U., Chamola, V., Benslimane, A., Guizani, M.: Toward safer vehicular transit: implementing deep learning on single channel EEG systems for microsleep detection. IEEE Transactions on Intelligent Transportation Systems (2021). https://doi.org/10.1109/TITS.2021.3125126
7. Malafeev, A., Hertig-Godeschalk, A., Schreier, D.R., Skorucak, J., Mathis, J., Achermann, P.: Automatic detection of microsleep episodes with deep learning. Frontiers in Neuroscience **15** (2021). https://doi.org/10.3389/fnins.2021.564098
8. Anonymous: Title anonymized for double-blind review. Under review (2025).
9. Shoorangiz, R., Weddell, S.J., Jones, R.D.: Prediction of microsleeps from EEG: preliminary results. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4650–4653 (2016). https://doi.org/10.1109/EMBC.2016.7591764
10. Baseer, A., Weddell, S.J., Jones, R.D.: Prediction of microsleeps using pairwise joint entropy and mutual information between EEG channels. In: Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4495–4498 (2017). https://doi.org/10.1109/EMBC.2017.8037855
11. Anonymous: Title anonymized for double-blind review. In: 33rd International Conference on Artificial Intelligence and Cognitive Science (AICS'25), Dublin (2026). DOI omitted for double-blind review.
12. Anonymous: Title anonymized for double-blind review. In: 33rd International Conference on Artificial Intelligence and Cognitive Science (AICS'25), Dublin (2026). DOI omitted for double-blind review.
13. Shoorangiz, R., Buriro, A.B., Weddell, S.J., Jones, R.D.: Detection and prediction of microsleeps from EEG using spatio-temporal patterns. In: Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 522–525 (2019). https://doi.org/10.1109/EMBC.2019.8857962
14. She, Q., Chen, K., Luo, Z., Nguyen, T., Potter, T., Zhang, Y.: Double-criteria active learning for multiclass brain-computer interfaces. Computational Intelligence and Neuroscience **2020**, 3287589 (2020). https://doi.org/10.1155/2020/3287589
15. Wu, D.: Active semi-supervised transfer learning (ASTL) for offline BCI calibration. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC) (2017). arXiv:1805.05781
16. Li, X., Hong, L., Wang, J.-C., Liu, X.: Fatigue driving detection model based on multi-feature fusion and semi-supervised active learning. IET Intelligent Transport Systems **13**(9), 1401–1409 (2019). https://doi.org/10.1049/iet-its.2018.5590
17. Cardoso, T.N.C., Silva, R.M., Canuto, S., Moro, M.M., Gonçalves, M.A.: Ranked batch-mode active learning. Information Sciences **379**, 313–337 (2017). https://doi.org/10.1016/j.ins.2016.10.037
18. Hertig-Godeschalk, A., Skorucak, J., Malafeev, A., Achermann, P., Mathis, J., Schreier, D.R.: Maintenance of Wakefulness Test (MWT) Recordings (Version 1) [Data set]. Zenodo (2019). https://doi.org/10.5281/zenodo.3251716