Laughter Accelerometer-Based Detection in Natural Social Interactions

Investigating segmentation and inter-modality annotation strategies for wearable laughter detection

Luka Knezevic Orbovic[1]

Supervisor(s): Hayley Hung[1], Litian Li[1], Stephanie Tan[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to the EEMCS Faculty of Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Luka Knezevic Orbovic
Final project course: CSE3000 Research Project
Thesis committee: Hayley Hung, Litian Li, Stephanie Tan

Abstract

We propose a method for detecting laughter in spontaneous social interactions using chest-worn accelerometers. Our approach compares three segmentation strategies—padded, centered, different sliding windows sizes and evaluates annotation modalities: *No Audio*, *Only Audio*, and *With Audio*. Using time-domain features and Random Forests, we reach up to 0.962 macro F1-score.

Longer windows and multimodal annotations improve performance and generalizability. Key features include axis-wise means, derivatives, and inter-axis correlations. These results support the potential of motion-based laughter detection in privacy-sensitive environments, while highlighting the importance of segment and label design.

1 Introduction

Laughter is a rich social signal involved in emotion regulation, bonding, and group cohesion [4]. Detecting laughter automatically has become an important goal in affective computing. While audio-based systems achieve high accuracy in controlled settings, they suffer in real-world scenarios due to noise, reverberation, and privacy concerns [11].

Wearable sensors offer a robust and privacy-preserving alternative. Chest-mounted accelerometers can detect thoracic and respiratory vibrations during laughter, without recording audio or facial expressions. Prior work shows this method works both in lab settings and spontaneous social interactions [6]. However, most studies rely on fixed-length windows, which may split laughter events or capture unrelated motion, reducing accuracy.

Laughter annotation adds further complexity. Annotators may use video only (No Audio), sound only (Only Audio), or both (With Audio), leading to inconsistent labels depending on the sensory input. Combining these labels into a single ground truth can further distort the concept of laughter [9].

These segmentation and annotation choices influence model performance and must be studied in interaction. For example, a segmentation strategy misaligned with how annotators perceived laughter may harm generalization. Thus, we explore the interplay between segmentation, feature extraction, and label construction in the laughter detection pipeline.

Our central research question is:

How effectively can chest-worn accelerometer data be used to detect laughter in spontaneous social interactions?

To answer this, we address four subquestions:

- How do contiguous segments compare to fixed-size windows for detecting laughter, and what duration yields the best performance?
- How do annotation strategies affect the performance and generalizability of models across inter-modality conditions?

- How does segmentation type (contiguous, sliding, or centered) impact model robustness?
- Which time-domain features carry the most predictive value for laughter detection from accelerometer data?

By aligning segments with behavioral boundaries and using interpretable features, our method supports privacy-aware laughter detection in naturalistic environments.

2 Related Work

Laughter detection has been widely explored, particularly in audio and audio–visual domains. Early systems relied on hand-crafted acoustic features such as pitch, energy, and spectral descriptors extracted from short fixed-length windows [11]. With the rise of deep learning, convolutional and recurrent models operating on Mel-spectrograms have significantly improved performance in controlled settings [11].

In privacy-sensitive and noisy environments, wearable sensors offer an alternative modality. Chest-worn accelerometers capture mechanical chest-wall vibrations correlated with laughter and respiratory patterns. Di Lascio et al. [6] used 5-second windows with time- and frequency-domain features to classify laughter, achieving approximately 81 percent accuracy in a lab-based study.

Segmentation strategies critically impact performance. Sliding-window approaches have been standard due to simplicity [3], but arbitrary window boundaries can dilute event-specific characteristics. In related domains, contiguous-event segmentation has shown advantages: cough detection using acceleration data segmented by burst onset/offset outperformed fixed windows [7], and gesture recognition benefited from run-length encoding of annotated frames [3].

Annotation strategy plays a critical role in laughter detection. Multi-annotator consolidation methods, such as majority voting, are commonly used in affective computing but may suppress rare or ambiguous events. In contrast, union voting emphasizes recall and has been employed in emotion recognition to better capture subtle expressions [9]. Annotation consistency and inter-rater agreement also depend heavily on the modality used. Providing annotators with synchronized audio previews for ambiguous accelerometer segments has been shown to boost agreement by up to 20 percentage points [9]. Each annotation modality reflects a distinct sensory input: No Audio (video-only), Only Audio (audio-only), and With Audio (audio and video). These modalities fundamentally influence how laughter is perceived. For instance, No Audio often misses vocal or subtle laughter, while Only

Audio may overlabel events lacking visual confirmation. The With Audio condition generally yields more consistent and temporally precise labels [9].

Feature engineering for wearable sensors encompasses both time- and frequency-domain representations. Signal magnitude area (SMA) and inter-axis correlations have proven effective for general activity recognition [1].

Classification algorithms range from traditional random forests and support vector machines [2] to ensemble and deep architectures [11]. Random forests remain popular in wearable analytics due to robustness and feature-importance interpretability [2].

3 Methodology

We designed a complete data pipeline for detecting laughter from chest-mounted accelerometer signals recorded in spontaneous social interactions. The process includes signal preprocessing, annotation handling, segmentation, padding, time-domain feature extraction, and Random Forest classification.

3.1 Dataset and Annotation

We use the ConfLab dataset [10], which features 48 participants engaged in spontaneous social interactions while wearing lanyard-mounted 3-axis accelerometers sampled at 56 Hz. Laughter events were annotated frame by frame at 60 Hz by three independent annotators under three different perceptual conditions—*No_Audio*, *Only_Audio*, and *With_Audio*—each representing a different annotation modality.

3.1.1 Modalities

Each annotation modality (*No_Audio*, *Only_Audio*, *With_Audio*) was treated as a separate labeling condition to preserve its perceptual perspective. We did not fuse modalities but instead used cross-modality evaluation to assess generalization beyond the original annotation context.

Within each modality, we combined the three annotators' frame-level labels using majority voting. This choice reduced false positives and eliminated unrealistically long laughter durations. For instance, union voting produced segments exceeding 200 seconds—far beyond typical laughter episodes, which usually last under 10 seconds [8]. Majority voting brought maximum segment lengths down to around 11 seconds, better aligning with known laughter dynamics.

3.1.2 Segmentation Strategy

We segment the accelerometer signal into contiguous regions of consistent annotation—laughter (label = 1) and non-laughter (label = 0)—to preserve the natural structure of the behavior. This strategy allows us to recover the full temporal extent of each episode without arbitrary cuts, maintaining alignment between the signal and the underlying event. It also enables us to compute meaningful statistics over event durations, such as the average and maximum length of laughter and non-laughter segments (Figure 1).

This is especially important in laughter detection, where episodes are typically brief and burst-like. In our dataset, most laughter events last under one second (median = 0.67s, mean = 0.97s), reinforcing the need to preserve fine-grained temporal boundaries. By grouping consecutive frames with the same label into a single segment, we ensure each window represents a coherent perceptual unit. To reduce label noise, we discard segments shorter than 200ms, as such brief instances are unlikely to reflect meaningful laughter events. Prior research shows that genuine laughter typically unfolds over longer intervals, both acoustically and behaviorally [8].

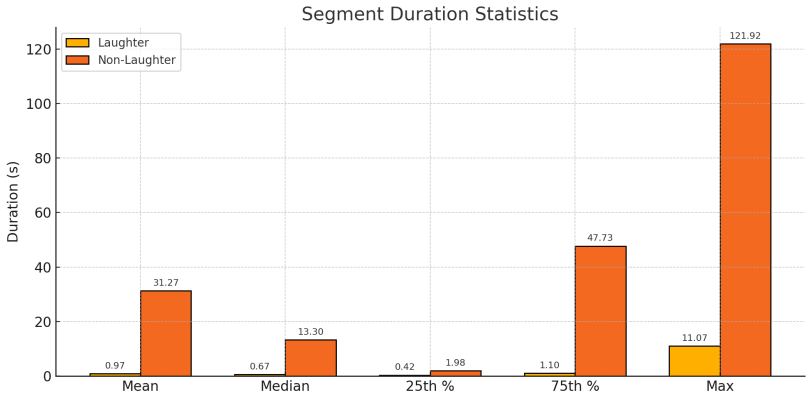


Figure 1: Duration statistics for contiguous laughter and non-laughter segments after filtering.

3.1.3 Padding Techniques

To prepare the data for statistical feature extraction and avoid biases introduced by variable-length segments, we restructured all contiguous regions into fixed-length windows of 1 s, 2 s, and 10 s. Most laughter events were short, with a mean duration of 0.97 s, making 1 s a natural fit a window size. The primary motivation for introducing longer windows was to preserve more usable laughter data. By including 2 s (around 90 percent) and 10 s windows (around 100 percent), we are able to retain more laughter examples in their entirety, improving dataset coverage at the cost of additional

padding. Additionally, we use these larger windows to explore whether segment length alone—independent of any temporal modeling—affects feature stability or classification performance. While longer windows may dilute the laughter signal when it occupies only a small portion of the segment, they also offer a broader context for statistical aggregation and may enhance discriminative power in some cases.

To prepare the accelerometer signal for classification, we explored three segmentation strategies: **padded segment blocks**, **timeline-centered windows**, which both come from continuous segments, and **fixed-size sliding windows**, all tested in all three window sizes:

1. Padded Segments

This strategy starts from the annotated laughter or non-laughter segments and adjusts their length to a fixed window size. If a segment is shorter than the target duration, we pad it by repeating the last value along each axis to avoid sharp discontinuities. If it is longer, we split it into multiple equal-sized chunks, padding the final one if needed. This ensures that all training inputs have uniform length, while staying tightly aligned with the temporal boundaries provided by the annotation. Padding with repeated values is well suited to our statistical time-domain features (e.g., mean, variance, energy), which are less sensitive to such repetition than frequency-based features [3]. A padded segment is considered positive if the continuous segment was considered laughter, as can be deducted.

2. Timeline-Centered Windows

Here, instead of modifying the segment length, we use it to determine where to extract a fixed-size window from the full accelerometer signal. Specifically, we center the window around the temporal midpoint of each annotated event and extract the desired duration from the surrounding signal. This means the window may include signal outside the continuous annotation, especially when segments are short. No padding is applied, and the extracted window reflects real signal values, including potentially noisy or ambiguous transitions. This approach better simulates real-world conditions where boundaries are not always clean, and models must learn to focus on the most informative part of the signal. A timeline-centered segment is considered laughter if the continuous segment was considered laughter.

3. Fixed-Size Sliding Windows

We also applied a traditional fixed-size sliding window approach, widely used in activity recognition for its simplicity and consistency. While it may not align with behavioral boundaries and can split short events like laughter,

it remains a strong baseline. We used window lengths of 1, 2, and 10 seconds with a 0.5-second overlap. A window was labeled as laughter if at least 10 percent of its frames were annotated as such—a threshold chosen to balance capturing enough laughter signal while minimizing noise. This is important to note: determining how much of a segment must contain laughter to justify a positive label is not trivial; this thresholding decision is an open research question with implications for both label quality and model sensitivity. To address class imbalance, the resulting dataset was resampled to maintain a 10–90 positive-to-negative ratio as originally the.

3.2 Signal Preprocessing

Raw accelerometer signals were obtained from 9-axis IMUs worn by each of the 48 participants at a sampling rate of 56 Hz. Following established preprocessing practices [5], the signals underwent the following steps:

- **Band-pass filtering (0.5–10 Hz)**: Applied a 4th-order Butterworth filter to isolate laughter-related chest vibrations, effectively removing low-frequency signals related to posture shifts or respiration and high-frequency noise such as tremors and motion artifacts.
- **Z-score normalization**: Standardized each accelerometer axis (accelX_filtered, accelY_filtered, accelZ_filtered) on a per-participant basis to account for inter-subject variability in resting posture and sensor orientation.

The filtered and standardized signals were serialized into `.pkl` files for efficient retrieval and reuse in subsequent analysis pipeline stages.

3.3 Feature Extraction

We extracted time-domain features from fixed-length segments of the standardized accelerometer signals (accelX_filtered, accelY_filtered, accelZ_filtered), sampled at 56 Hz. These features follow prior work in physiological state modeling with inertial sensors [3].

For each axis (X, Y, Z), we computed:

- **Mean**, **Variance**, **Energy**:

$$\text{Energy} = \sum_{t=1}^{N} x_t^2$$

- **Mean** and **Standard deviation of the first derivative**

Additionally, we included:

- **Signal Magnitude Area (SMA)** across all axes:

$$\text{SMA} = \sum_{t=1}^{N} (|X_t| + |Y_t| + |Z_t|)$$

- **Inter-axis correlations** (`corr_xy`, `corr_xz`, `corr_yz`), via Pearson's coefficient.

This results in a 19-dimensional feature vector per segment: 15 from axis-specific statistics (5 per axis), plus SMA and 3 inter-axis correlations.

3.4 Random Forest Classification

We trained a Random Forest (RF) classifier from scikit-learn to distinguish laughter from non-laughter segments based on the extracted features. The RF classifier was chosen for its proven effectiveness in prior laughter detection studies, robustness to varying scales, efficiency with limited datasets, and interpretability through feature importance analysis. The following protocol was employed:

- **Participant-Disjoint Splits**: A 75-25 training-test split was applied specifically to sliding window datasets, ensuring no participant appeared in both sets, thus preventing identity leakage and enhancing generalization.
- **Cross-Modality and Segmentation Testing**: Since evaluations included different modalities and segmentation strategies, the participant-disjoint split was explicitly applied only when testing within the same dataset.
- **Hyperparameter Tuning**: Optimized using Weights and Biases (wandb), exploring parameters such as tree number and maximum depth.

Each model was systematically trained and evaluated for every annotation modality, window size (1s, 2s, 10 s), and segmentation strategy (padded, centered, sliding). We reported average accuracy, F1-score, and feature importance based on mean decrease in Gini impurity, explicitly addressing our research questions regarding optimal segment duration, annotation strategy generalizability, and predictive feature identification.

4 Results

We evaluate our Random Forest classifier for each window size, each modality, and each segmentation strategy. We report precision, recall, and F1-score for both laughter and non-laughter classes.

We structure our evaluation into three setups: **intra-modality** (training and testing on the same annotation modality), **inter-modality** (training on one modality and testing on another), and **inter-segmentation** (training on continous segments and then testing on a spontaneous real life encounters).

4.1 Intra-Modality Performance

To evaluate the effectiveness of our approach, we trained separate Random Forest classifiers for each annotation modality—*No Audio*, *Only Audio*, and *With Audio*—across three window sizes: 1 s, 2 s, and 10 s, under three segmentation strategies: *centered*, *padded*, and *sliding*. For each configuration, we evaluated the macro-averaged F1-score ($\mathrm{F1}_{\mathrm{macro}}$), as well as class-specific metrics for laughter ($\mathrm{F1}_1$, Recall_1).

Figure 2 summarizes the performance across all configurations, with a separate subplot per modality. Each line represents a segmentation strategy, plotted across increasing window sizes.

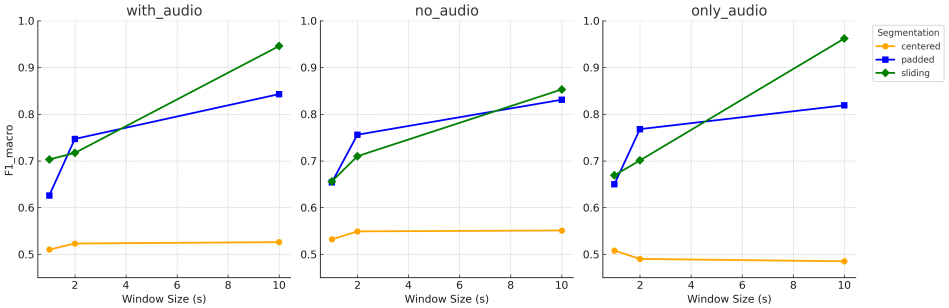


Figure 2: Macro-averaged F1-score across window sizes and segmentation strategies, trained and tested on the same modality.

We observe that the *sliding window* strategy consistently achieves the highest performance across all modalities, particularly at the 10-second window. This suggests that overlapping windows provide greater temporal context, enabling the model to detect laughter with higher reliability.

The *padded* strategy also performs competitively, especially in the 10-second condition, confirming that aligning segment lengths with annotated laughter intervals benefits statistical feature extraction. In contrast, the *centered* approach yields lower F1 scores, likely due to limited contextual information and potential misalignment with the laughter event.

Overall, longer and overlapping segments (10 s sliding) proved most effective, supporting the hypothesis that laughter detection benefits from broader temporal integration. The differences between segmentation strategies further emphasize the importance of segment design in modeling spontaneous social behaviors.

For completeness, detailed bar plots showing $\mathrm{F1}_{\mathrm{macro}}$, $\mathrm{Precision}_1$, and Recall_1 for each modality and segmentation type are provided in Appendix A.

4.2 Inter-Modality Performance

We conducted an inter-modality generalization analysis by training a Random Forest classifier on features extracted from one annotation modality and testing on another. This setup is crucial to assess the robustness of models across annotation perspectives—e.g., visual-only (*No Audio*), acoustic-only (*Only Audio*), and combined cues (*With Audio*).

We evaluated all train-test combinations across three segmentation strategies—*padded*, *centered*, and *sliding*—and three window durations: 1 s, 2 s, and 10 s. For each configuration, we computed macro-averaged F1-score ($\mathrm{F1}_{\mathrm{macro}}$) and report the results grouped by train-test pair.

Figure 3 provides a visual summary of this evaluation. Each subplot corresponds to a different train→test condition, with separate lines for each segmentation strategy plotted across increasing window sizes.

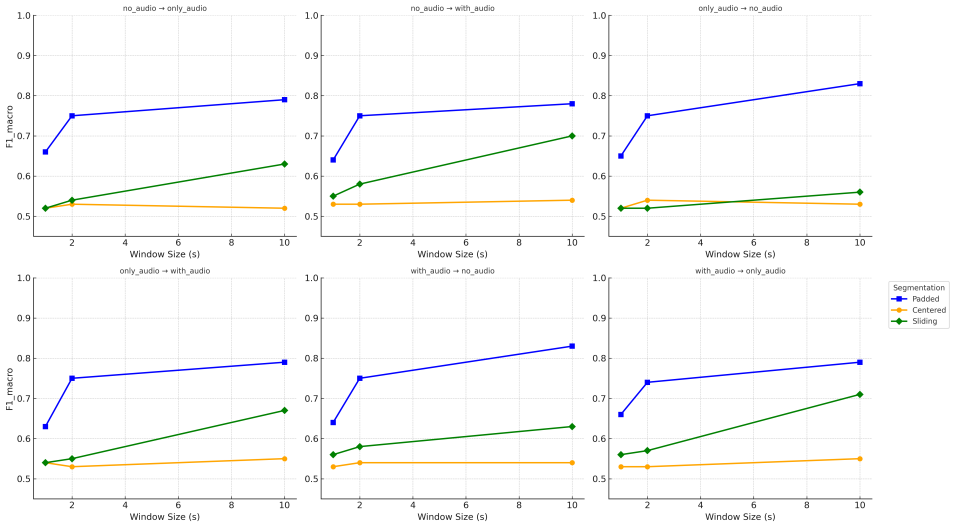


Figure 3: Inter-modality macro-averaged F1-score across window sizes, segmentation strategies, and modalities grouped by train→test condition.

We observe that longer window sizes (10 s) consistently led to improved cross-modal generalization, particularly under the *padded* and *sliding* segmentation strategies. Sliding windows often yielded the strongest generalization performance, especially when transferring from `with_audio` annotations, which provide richer supervision due to the inclusion of both visual and auditory cues.

In contrast, the *centered* segmentation strategy produced relatively flat performance across durations and often underperformed in transfer settings,

suggesting that limited temporal context harms generalization.

These results reinforce the importance of both window size and segment alignment strategy in building robust models for spontaneous behavior recognition. Full precision results and bar plots per condition are included in Appendix B.

4.3 Inter-Segmentation Performance

We evaluated the two continuous segmentation strategies—*padded* and *centered*—to test their robustness when applied to the sliding window framework.

As shown in Figure 4, neither segmentation strategy achieved consistently satisfactory results. Macro F1-scores rarely exceeded 0.5, and recall was generally low across modalities and window sizes. The padded strategy yielded higher accuracy overall, but this came at the expense of extremely low recall—particularly for non-audio modalities—revealing a strong bias toward predicting the dominant non-laughter class. Centered segmentation slightly improved this trade-off by capturing better recall, but still suffered from imbalanced precision and limited overall robustness.

While average performance was modest, certain modality-duration combinations stood out. Most notably, the *with_audio* modality under the padded strategy achieved a remarkably high precision_1 of 0.83 with 10-second windows. This indicates that when the model predicted laughter, it was often correct—even if such predictions were rare. However, this came with extremely low recall, suggesting a highly conservative classifier that may only respond to more exaggerated or unambiguous laughter cues. These modality-specific outliers highlight the importance of context: multimodal annotations paired with longer input windows can lead to more confident, albeit sparse, detections.

These findings suggest that behavioral alignment alone—whether padded or centered—is not sufficient to overcome the inherent challenges of detecting laughter from raw accelerometer signals. Segment duration and modality both play important roles in shaping model sensitivity and bias. A more detailed performance breakdown per segmentation and modality—including precision_1 and recall_1—is available in Appendix B.4.

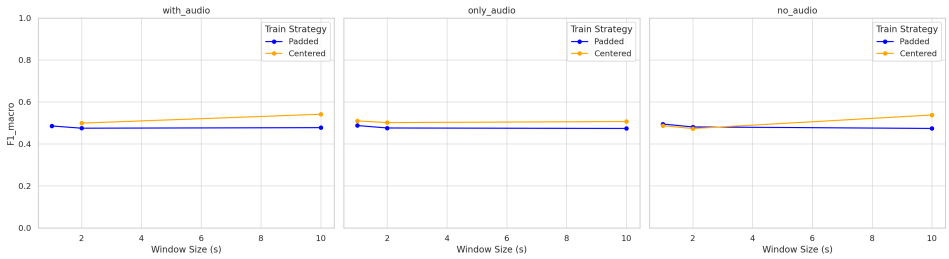


Figure 4: F1-macro scores for each sampling strategy (padded and centered) on sliding windows.

4.4 Feature Importance

To interpret the decision patterns of our Random Forest classifiers, we analyzed feature importances—measured via Gini importance—from the best-performing models at each window size (1 s, 2 s, 10 s), across the three segmentation strategies. Appendix C presents the corresponding plots.

In the *padded* condition, the classifier strongly favored mean-based features—particularly accelY_filtered_mean, accelX_filtered_mean, and accelZ_filtered_mean—across all window sizes. These features capture overall postural shifts, suggesting that laughter events tend to modulate average chest movement. Derivative statistics and cross-axis correlations such as corr_yz and corr_xy also ranked consistently, indicating their value in capturing short-term coordination patterns during laughter.

In the *event-centered* strategy, a similar feature ranking emerged, but with a slightly stronger emphasis on derivative-based features such as accelX_filtered_deriv_std and accelZ_filtered_deriv_std. These highlight intra-laughter movement variation that is well captured when windows are precisely centered around the laughter segment. The use of SMA and axis-specific variances suggests the model also benefits from integrating signal energy and dispersion in these temporally aligned windows.

Under the *sliding window* approach, importance shifted toward derivative and variance-based descriptors, particularly in the 10-second window condition. Features like accelX_filtered_deriv_std and accelY_filtered_deriv_std dominated across durations, reflecting the method's reliance on broader temporal integration to capture patterns that may span beyond laughter boundaries. Compared to the other strategies, sliding windows favored dynamic signal statistics over static posture descriptors.

Across strategies, features like axis-specific means, derivatives, and cross-axis correlations carried the most predictive value. Their prominence confirms that laughter involves both posture shifts and dynamic motion signatures, aligning with prior work on thoracic movement patterns.

5 Responsible Research

5.1 Ethical Considerations

This project involved analyzing laughter behavior using accelerometer data collected during spontaneous social interactions. All data used in this research comes from the Conflab dataset [10], which was ethically approved and collected in accordance with TU Delft's human research ethics guidelines. Participants gave informed consent, and all data was anonymized prior to release. No audio or video recordings were used directly; our analysis relied exclusively on inertial motion data and externally provided annotations, ensuring participant privacy throughout the study.

Furthermore, this work does not attempt to infer or identify personal traits, emotions, or identities beyond detecting the presence of laughter. We intentionally avoided training models for individual classification or demographic prediction, as such uses could lead to ethically questionable applications. Instead, our focus was strictly on behavioral event detection using non-invasive, privacy-preserving signals.

5.2 Reproducibility and Research Integrity

To ensure reproducibility and scientific transparency, all components of our pipeline—from signal preprocessing and annotation handling to segmentation, feature extraction, and classification—were implemented in Python using publicly available libraries such as `scikit-learn`, `NumPy`, and `pandas`. Signal processing and filtering routines followed established practices and are clearly documented.

The project was developed and executed in Jupyter notebooks, with core functionality modularized into reusable classes. Annotated and filtered signals were stored in serialized `.pkl` files, and all parameters (e.g., window sizes, segmentation types, filter cutoff frequencies) are explicitly defined. Random Forest hyperparameters were tuned using Weights & Biases (wandb), and all experiment runs were tracked with fixed seeds to support consistent replication.

The full codebase, along with sample data loaders, annotation preprocessing scripts, and experiment logs, will be made available upon request or publication. Our evaluation protocol used participant-disjoint splits to prevent data leakage, and all inter-modality experiments were performed with rigorous cross-testing procedures to avoid overfitting. Through these design choices, we aim to promote reproducible, interpretable, and ethically sound research in wearable affective computing.

5.3 Use of Language Models:

Large Language Models (LLMs) such as ChatGPT were used throughout the project to assist with code debugging, data analysis brainstorming, and editing of scientific writing. All outputs from the model were critically reviewed, validated, and edited by the researcher to ensure accuracy, originality, and alignment with scientific standards. LLMs were treated as productivity-enhancing tools rather than content generators, and no unverified model output was included without modification or scrutiny.

6 Discussion

This section reflects on the implications of our findings within the broader context of affective computing and wearable sensing. We present four central discussion points—each addressing a key methodological or empirical insight—followed by reflections on how these findings inform future research directions.

6.1 Model Performance and Interpretability

Despite their simplicity, Random Forest classifiers achieved robust performance across most configurations. This indicates that laughter generates repeatable, structured motion patterns that can be captured using statistical features alone. The models' reliance on time-domain features—such as axis-specific means, variance, derivatives, and inter-axis correlations—provides physiological interpretability, linking detection performance to respiratory and thoracic movements.

However, this reliance also limits the model's ability to capture temporal dynamics. Without access to sequential modeling or frequency-domain features, the classifiers miss expressive characteristics like rhythmic oscillations or buildup. This was particularly evident when testing on different segmentation strategies, where broader temporal integration often diluted segment-specific information and reduced performance sharply.

6.2 Annotation Modality and Label Consistency

Training and testing across modalities highlighted the impact of perceptual bias in laughter annotation. Models trained on multimodal annotations (i.e., With Audio) consistently outperformed those trained on video-only or audio-only labels, particularly when tested on different modalities. This suggests that annotations grounded in richer perceptual information (both audio and visual) produce more generalizable models.

Conversely, models trained on No Audio annotations showed weaker transfer performance, likely due to their overemphasis on exaggerated visual

cues that don't always align with acoustic laughter. This reflects a key limitation: while annotation modality affects model performance, the dataset offers no definitive ground truth. The variation among human annotators across modalities reveals how subjective the definition of laughter can be, which challenges efforts to design consistent affective computing systems.

6.3 Segmentation Strategy and Temporal Scope

Our findings highlight a trade-off between segment duration and ecological validity. Although models trained on 10-second segments consistently outperformed those using 1- or 2-second windows, most laughter events in the dataset were much shorter. The padded segmentation strategy extends short events by repeating signal content to reach fixed durations, which improves performance. However, this comes at a cost: artificial padding may distort the temporal dynamics of laughter and delay response in real-time applications.

Alternative strategies present different trade-offs. Centered windows require knowledge of laughter midpoints, which is impractical at runtime. Sliding windows, by contrast, are runtime-friendly and performed best overall—particularly with longer durations—thanks to their broader temporal coverage. While they may occasionally dilute label quality by overlapping non-laughter motion, their consistent performance highlights the value of overlapping, context-rich segments.

6.4 Robustness Across Temporal and Perceptual Conditions

We assessed model performance across annotation modalities and segmentation strategies. Stronger generalization occurred when training on well-aligned segments, such as 10-second padded windows or centered windows with *With Audio* labels.

Despite these gains, reliance on fixed-size windows and same-modality training limits broader deployment. Future models should adapt to noisy labels and dynamic timing in real-world settings. This could be done by testing more the robustness of sliding windows in laughter detection, its ideal positive label interpretation, or resampling techniques, also enhance model expressiveness through temporal architectures (e.g., LSTMs, TCNs) or spectral features.

7 Conclusion

This project investigated the effectiveness of chest-worn accelerometers for detecting laughter in spontaneous social interactions, focusing on the interplay between segmentation strategies, annotation modalities, and classifier robustness. Our results show that even with a non-sequential model like a

Random Forest, careful pipeline design can achieve high performance, with macro F1-scores reaching 0.962 (Only Audio, 10s sliding windows).

We now revisit the central research question:

How effectively can chest-worn accelerometer data be used to detect laughter in spontaneous social interactions?

Accelerometer data—when paired with appropriate segmentation, label consolidation, and feature extraction—can reliably detect laughter while preserving privacy. Our findings confirm the viability of this approach in naturalistic environments.

Fixed-size windows, particularly 10-second sliding segments, consistently outperformed behaviorally aligned contiguous ones. Although contiguous segments reflect natural event boundaries, they introduced class imbalance and label noise that hindered performance. Sliding windows, by contrast, provided broader coverage and more diverse samples, which improved overall detection—even when windows contained mixed content.

Annotation modality also played a key role in model generalization. Models trained on *With Audio* labels achieved the best cross-modality results, likely due to richer perceptual cues combining visual and auditory signals. In contrast, *No Audio* labels resulted in the weakest transfer, suggesting over-reliance on visual exaggerations and overall lower label quality.

Among segmentation types, sliding windows delivered the most robust and consistent classification results across durations and modalities. Padded segments performed competitively when tightly aligned to laughter episodes, while centered windows, despite being precisely located, lacked contextual information and generalized poorly.

Across all conditions, the most predictive features included axis-wise means and derivative-based statistics (e.g., accelY_filtered_mean,

accelX_filtered_deriv_std)—capturing both postural changes and motion dynamics. Features like cross-axis correlations and signal magnitude area (SMA) also contributed, highlighting the high-energy and coordinated nature of laughter events.

In summary, our study supports the use of wearable accelerometers for laughter detection in real-world conditions. By addressing segmentation fidelity, annotation subjectivity, and feature design, we provide a foundation for privacy-aware and socially perceptive systems. Future work could explore sequential models or real-time applications to further improve responsiveness and reduce false positives.

References

[1] Limin Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. In *Proceedings of Pervasive Computing*, pages 1–17. Springer, 2004.

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*, 46(3):33:1–33:33, 2014.

[4] Robin I. M. Dunbar. Laughter and its role in the evolution of human social bonding. *Philosophical Transactions of the Royal Society B*, 377:20210176, 2022.

[5] Hayley Hung, Gwenn Englebienne, and Jeroen Kools. Classifying social actions with a single accelerometer. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 207–210, Zurich, Switzerland, 2013. ACM.

[6] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. Laughter recognition using non-invasive wearable devices. In *Proceedings of the 13th EAI International Conference on Pervasive Healthcare Technologies (PervasiveHealth '19)*, pages 262–271. ACM, 2019.

[7] Madhurananda Pahar, Igor Miranda, Andreas Diacon, and Thomas Niesler. Automatic non-invasive cough detection based on accelerometer and audio signals. *arXiv preprint arXiv:2109.00103*, 2021.

[8] Robert R. Provine. Laughing, tickling, and the evolution of speech. *Current Biology*, 14(6):R260–R261, 2004.

[9] Jose David Vargas Quiros, Laura Cabrera-Quiros, Catharine Oertel, and Hayley Hung. Impact of annotation modality on label quality and model performance in the automatic assessment of laughter in-the-wild. volume 15, pages 519–534, 2024.

[10] Chirag Raman, Jose Vargas-Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. Conflab: A data collection concept, dataset, and benchmark for machine analysis of free-standing social interactions in the wild. In *NeurIPS Datasets and Benchmarks Track*, 2022.

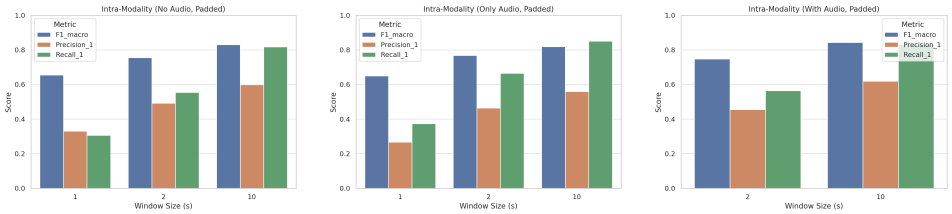
[11] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin,

Anna Polychroniou, Fabio Valente, and Samuel Kim. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Interspeech 2013*, pages 148–152, 2013.

Appendix

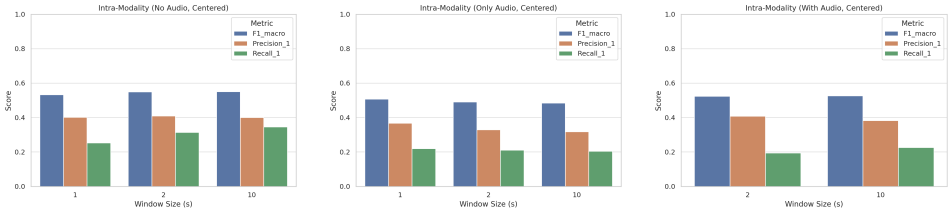
A Intra-Modality Performance Across Segmentations

This appendix includes full bar plot results for each segmentation strategy—*padded*, *event-centered*, and *sliding*—across all three annotation modalities and three window sizes (1s, 2s, 10s). Each figure shows macro F1, $F1_1$, and $Recall_1$ per modality.



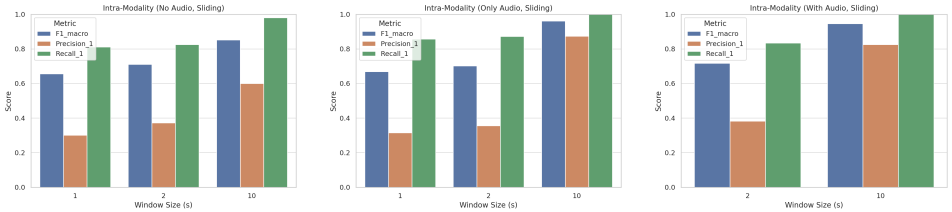
(a) No Audio modality (b) Only Audio modality (c) With Audio modality

Figure 5: Modality-wise performance using *padded* segmentation.



(a) No Audio modality (b) Only Audio modality (c) With Audio modality

Figure 6: Modality-wise performance using *event-centered* segmentation.



(a) No Audio modality (b) Only Audio modality (c) With Audio modality

Figure 7: Modality-wise performance using *sliding window* segmentation.

B Inter-Modality Performance Plots

We report detailed bar plots of inter-modality classification performance for all segmentation strategies (Padded, Sliding, Centered) and window sizes (1s, 2s, 10s). Each plot visualizes the F1 Macro, Precision_1, and Recall_1 scores when training on one annotation modality and testing on another. These results allow for a granular comparison of cross-modal generalization under varying temporal and segmentation assumptions.

B.1 Padded Segmentation

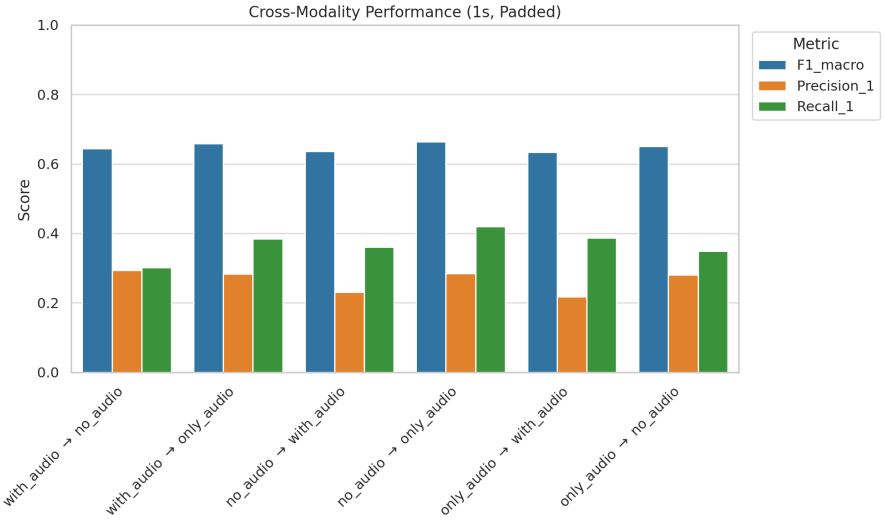


Figure 8: Inter-modality performance (1s windows, Padded)

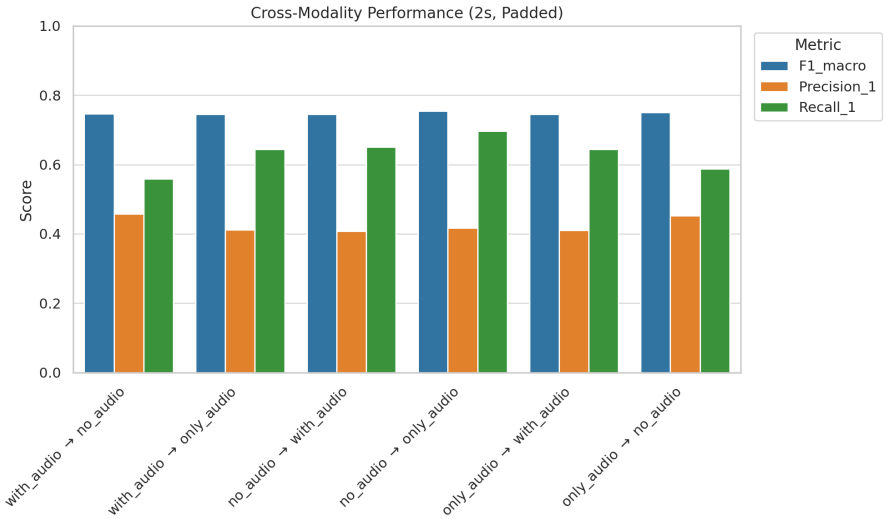


Figure 9: Inter-modality performance (2s windows, Padded)

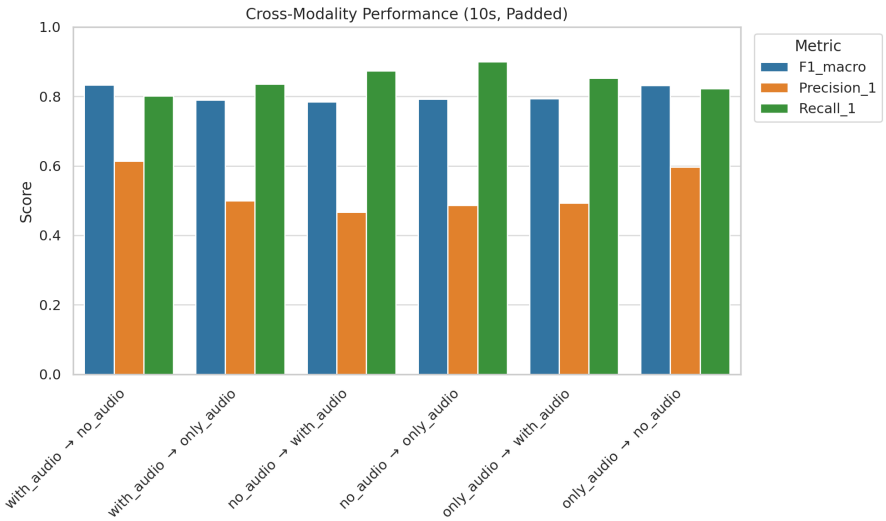


Figure 10: Inter-modality performance (10s windows, Padded)

B.2 Sliding Segmentation

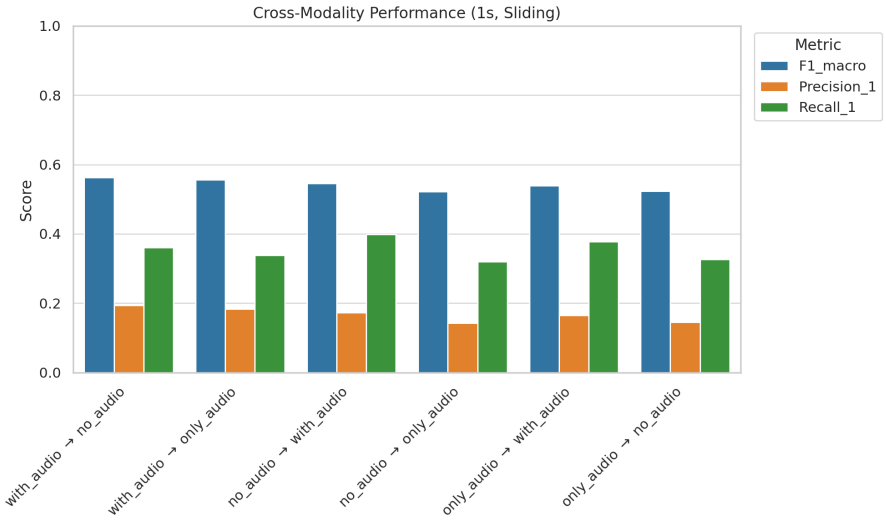


Figure 11: Inter-modality performance (1s windows, Sliding)

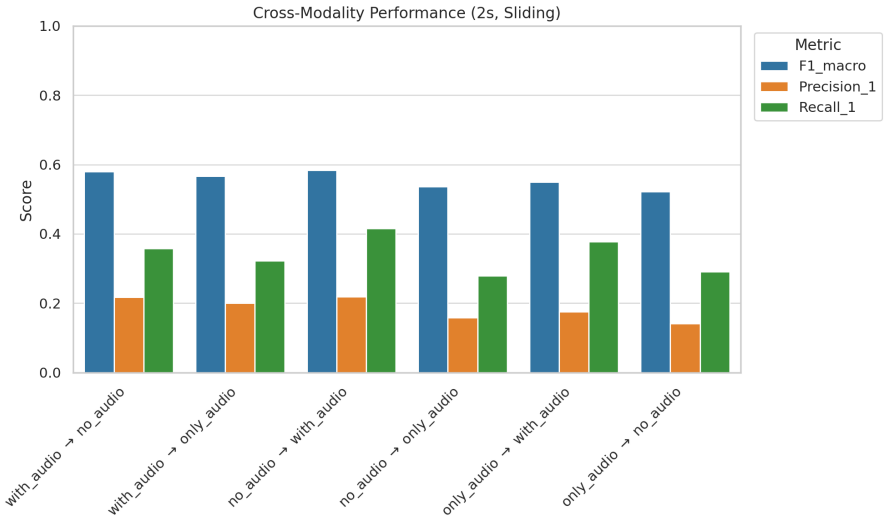


Figure 12: Inter-modality performance (2s windows, Sliding)

Figure 13: Inter-modality performance (10s windows, Sliding)

B.3 Centered Segmentation

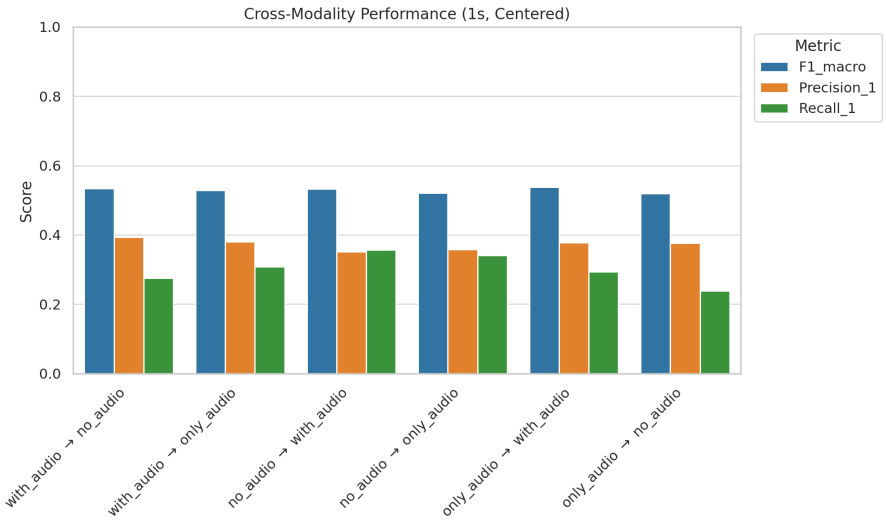


Figure 14: Inter-modality performance (1s windows, Centered)

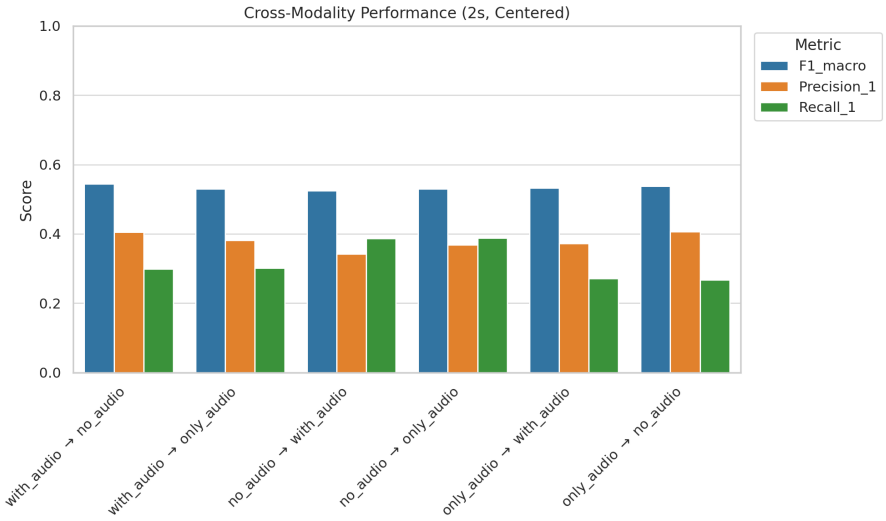


Figure 15: Inter-modality performance (2s windows, Centered)

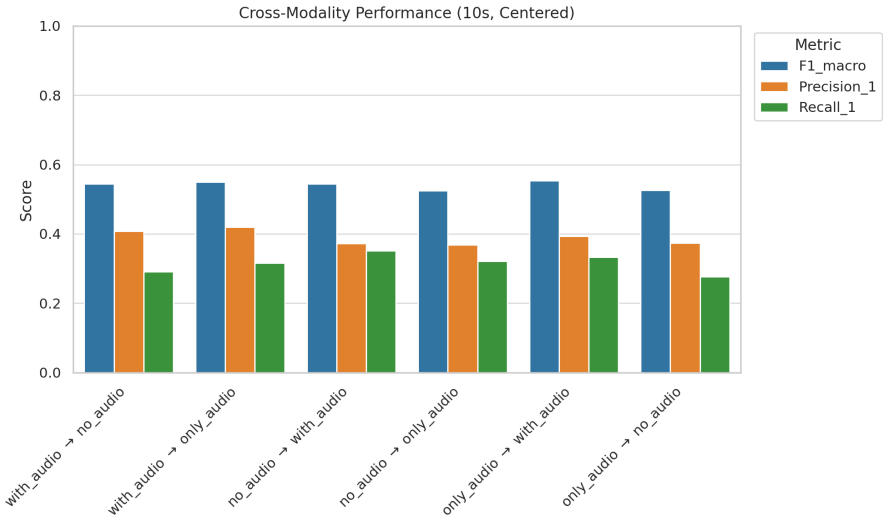


Figure 16: Inter-modality performance (10s windows, Centered)

B.4 Inter-Segmentation Performance by Modality and Duration

Figure 17 provides a detailed comparison of F1-macro, Precision, and Recall for each combination of segmentation strategy and modality, across all window sizes. Each subplot isolates one segmentation strategy, showing performance variations by modality and segment duration. These results further illustrate that padded segmentation can yield high precision (e.g.,

up to 0.83 for *with_audio*), but often at the cost of very low recall, especially in non-audio modalities.

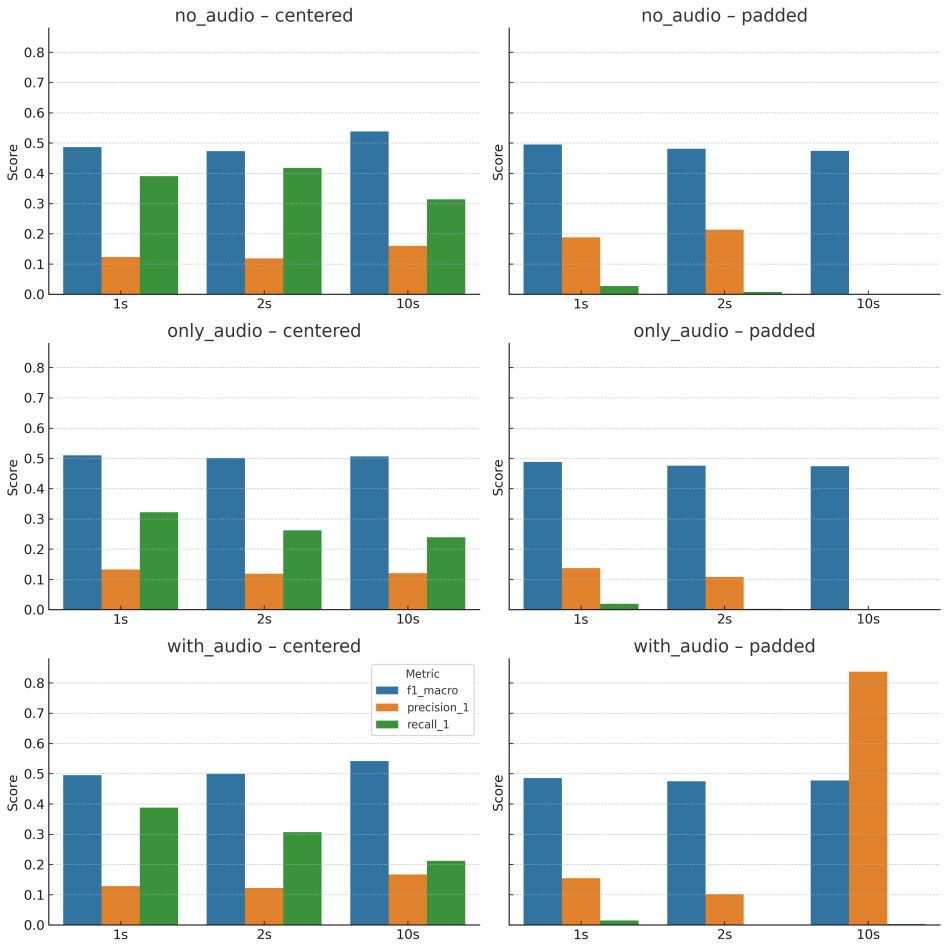
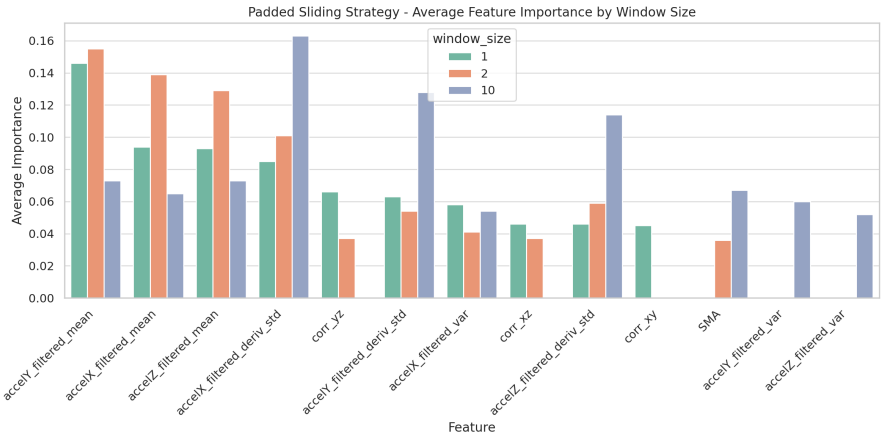


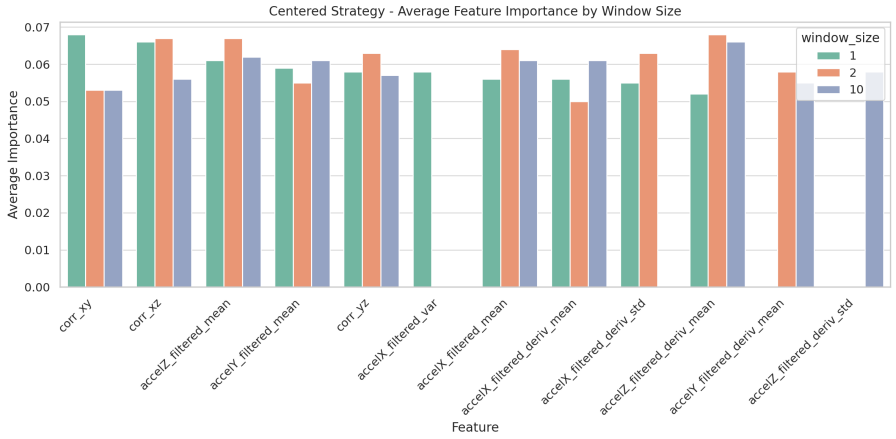
Figure 17: Performance scores (F1-macro, Precision, Recall) by modality and window size, split by segmentation strategy.

C Feature Importance Plots

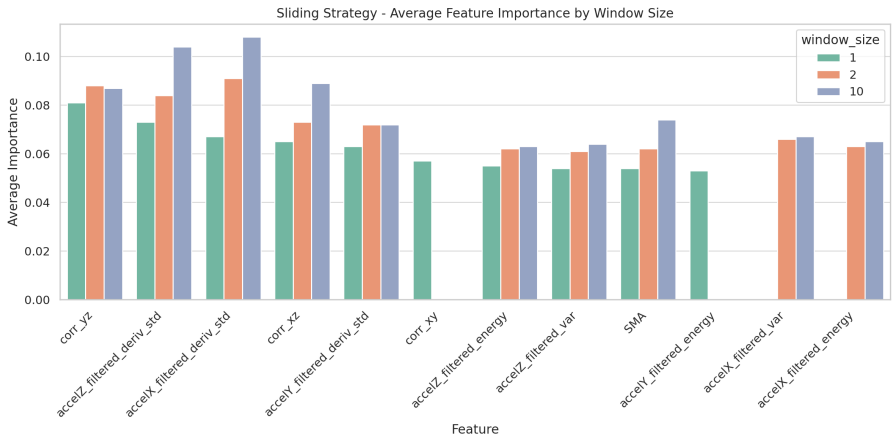
This appendix includes the detailed bar plots of feature importances for each segmentation strategy (padded, event-centered, and sliding), evaluated at all window sizes (1 s, 2 s, 10 s).



(a) Feature importances across window sizes (1 s, 2 s, 10 s) for padded segmentation.



(b) Feature importances across window sizes (1 s, 2 s, 10 s) for event-centered segmentation.



(c) Feature importances across window sizes (1 s, 2 s, 10 s) for sliding segmentation.

Figure 18: Feature importances from the best Random Forest model at each window size. Importance scores reflect the average Gini reduction across trees in the ensemble.