Multi-Modal Highlight Detection in Broadcast Audio: A Deep Learning Approach for Event Recognition in Sports and eSports

Nuno Costa[1][a], António Oliveira[1][b], Armindo Lobo[1][c], Ricardo Teixeira[1][d], Duarte Fernandes[1][e], Ricardo Rodrigues[1][f] and Emanuel Gouveia[2][g]

[1]*DTx – Digital Transformation CoLAB and the ALGORITMI Center, University of Minho, Guimarães, Portugal*

[2]*Universidade Católica Portuguesa, Instituto de Computação e Ciência de Dados, Portugal*

Keywords: AI-Based Sports Technologies, Machine Learning for Real-Time Analysis, Multi-Modal Deep Learning, Broadcast Stream Automation, Audio Event Detection.

Abstract: The detection of highlights in broadcast streams is essential for enhancing User Experience (UX) through automated summaries and efficient content retrieval. This is particularly relevant for live streaming environments common in sports and eSports, where audiences demand near real-time analysis. This paper presents a benchmark of models for highlight detection in broadcast audio, validated on the SoccerNet dataset but applicable to general competitive gaming streams. We propose a novel multi-modal architecture combining high-level semantic audio features (YAMNet) with Natural Language Processing (NLP) of transcribed commentary (analogous to eSports shoutcasting). Results show that fusing audio event detection with semantic text analysis significantly outperforms uni-modal baselines. The proposed framework offers a computationally efficient solution for AI-based broadcasting technologies, enabling scalable automation for content creators and improved viewer experiences.

1 INTRODUCTION

Highlight identification for broadcasts and streaming enables timely short-form content and improves engagement on fast-moving platforms. Because many sports and eSports are consumed live, near real-time identification of highlights enables broadcasters and streaming services to offer instant replays, short clips, and automated alerts, improving viewer experience even for audiences who cannot watch full events. Manual annotation is costly, automated detection scales these workflows and supports low-latency production.

Audio complements video by supplying emotional cues from crowd reactions and commentator emphasis, offering a lightweight signal for near real-time detection and mitigating some visual limitations (occlusions, replays, camera angles) (Vanderplaetse and Dupont, 2020; Nergård Rongved et al., 2021; Midoglu et al., 2024). Prior work used low-level acoustic descriptors and rule-based methods (Xiong et al., 2003; Baijal et al., 2015; Duxans et al., 2009; Harb and Chen, 2003; Dagtas and Abdel-Mottaleb, 2001), while newer approaches fuse audio and vision via deep models on log-mel spectrograms (Raventós et al., 2015; Vanderplaetse and Dupont, 2020; Nergård Rongved et al., 2021).

We propose representing audio at a higher semantic level by using selected class scores from a pretrained audio classifier together with a commentary-derived text score. This compact, interpretable feature set favors deployment (edge/low-latency) and generalizes better than many handcrafted feature pipelines. We note the approach is applicable beyond soccer (e.g., eSports) but requires domain validation and calibration for different broadcast styles.

Our implementation, developed with industry partners as a decision-support pipeline for audio analysis, uses public SoccerNet data for reproducibility (because confidentiality prevents releasing proprietary datasets) and relies on pretrained extractors (YAMNet for audio, Whisper for transcription) com-

[a] https://orcid.org/0000-0002-8425-3501
[b] https://orcid.org/0000-0002-2349-2373
[c] https://orcid.org/0000-0002-1517-9328
[d] https://orcid.org/0009-0004-5563-7719
[e] https://orcid.org/0000-0001-9736-5812
[f] https://orcid.org/0000-0001-7986-3754
[g] https://orcid.org/0000-0002-7785-2047

Costa, N., Oliveira, A., Lobo, A., Teixeira, R., Fernandes, D., Rodrigues, R. and Gouveia, E.
Multi-Modal Highlight Detection in Broadcast Audio: A Deep Learning Approach for Event Recognition in Sports and eSports.
DOI: 10.5220/0014585200004052

In *Proceedings of the 18th International Conference on Agents and Artificial Intelligence (ICAART 2026) - Volume 1*, pages 989-996
ISBN: 978-989-758-796-2; ISSN: 2184-433X

bined with lightweight supervised classifiers. Runtime profiling on representative hardware is planned as follow-up engineering work.

Paper structure: Section 2 reviews related work. Section 3 details our methodology and implementation. Section 4 reports experiments, Section 5 discusses results, and Section 6 concludes.

2 RELATED WORK

Audio-based highlight detection ranges from early energy- and repetition-based heuristics and handcrafted acoustic descriptors (Duxans et al., 2009; Rui et al., 2000; Dagtas and Abdel-Mottaleb, 2001; Harb and Chen, 2003; Baijal et al., 2015) to convolutional models on log-mel spectrograms (Simonyan and Zisserman, 2015; Vanderplaetse and Dupont, 2020; Nergård Rongved et al., 2021). Low-level features are lightweight and useful for low-latency pipelines but can suffer false positives and domain shift.

More recent work uses pretrained audio backbones and multimodal fusion to improve event spotting and robustness (Boes and Van hamme, 2021). Semantic audio classifiers such as YAMNet provide compact, interpretable representations useful for efficient downstream models and transfer learning (Gemmeke et al., 2017; Valliappan et al., 2024), while large transcription ASR systems (e.g., Whisper) supply complementary textual signals from commentary (Radford et al., 2023).

Multimodal approaches that combine audio semantic scores and commentary have shown improved highlight selection and generalization when telemetry is unavailable (Filippidis et al., 2018; Xiong and Wang, 2023; Ringer et al., 2022). Our method follows this trend: a small set of YAMNet class scores plus a lightweight commentary-derived feature are combined with a compact supervised classifier to produce interpretable, low-latency highlight candidates.

3 METHODS

3.1 Problem Characterization

Audio-based highlight detection must separate genuine excitement from routine crowd noise and cope with variability across matches and venues. The audio channel is inherently incomplete, so our method is intended as a lightweight, low-latency complement to video-based analysis rather than a replacement. Labels from SoccerNet are treated as noisy proxies of fan-interest, approximating what an average fan might consider interesting events, rather than representing a definitive identification of truly noteworthy moments.

3.2 Dataset Preparation

3.2.1 SoccerNet Dataset with Labeled Highlights

SoccerNet-v2 (henceforth referred to simply as SoccerNet) is a dataset with videos of 550 soccer game broadcasts (2 videos per game, 1 per half) from 6 major European soccer leagues. Among the intended tasks for the dataset is action spotting, for which 17 classes of events are temporally annotated by an anchor timestamp (Giancola et al., 2018) (Deliege et al., 2021). The broadcasts in SoccerNet contain narrations by commentators who mostly speak in 10 different languages, which were identified using Google Translate and the Whisper language detection model (English, Spanish, Russian, German, French, Turkish, Italian, Polish, Bosnian and Hungarian), except in 35 games where the language is not identifiable (Gautam et al., 2024).

3.2.2 Selection of Relevant Highlights

Considering the task of highlight detection, the 17 SoccerNet labels will be condensed into only 2, 'highlight' and 'non-highlight'. The events under the 'highlight' label are the ones that originally had the labels *Goal*, *Direct free-kick*, *Penalty*, *Shots on target*, *Yellow* $\rightarrow$ *red card*, *Red card* and *Yellow card*, while the rest is labeled as 'non-highlight'. These types of events were chosen as representative of highlights since they had greater effect on the outcome of the game, as well as higher correlation with fan interaction and possibly with sound activation patterns, making them the best candidates to be included as part of the automatic highlight detection.

3.3 Model Description

The proposed model architecture in Figure 1 has the following components: Feature Extractor — this initial stage processes the audio files to generate a set of relevant features (audio-processing node: YAMNet feature extraction and Whisper transcription) and the feature-integration node (which unifies features from the two models); Supervised Classifier — this modeling component takes the extracted features as input and projects them into a lower-dimensional subspace, outputting a probability score for each time

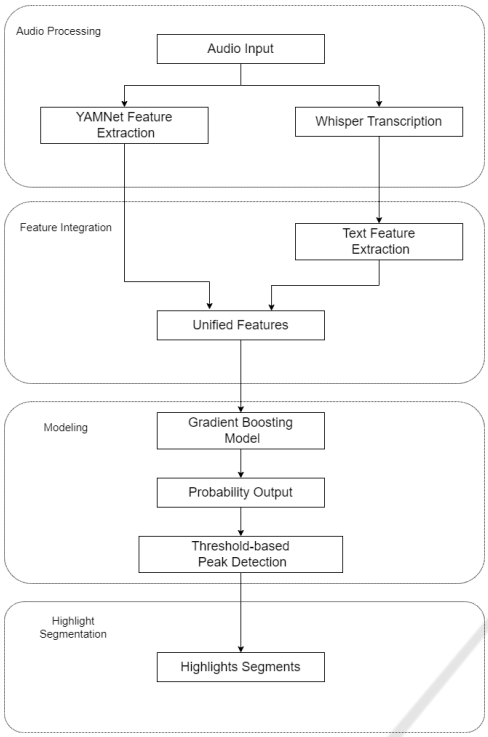

Figure 1: Overall architecture of our highlight detector.

point, which we refer to as the "signal"; Peak Detector (modeling) — this component processes the signal using threshold-based peak detection to identify intervals of high probability, which correspond to potential highlights and are then segmented (highlight-segmentation node).

We evaluated several options for feature extraction, classifier architectures, and peak detection algorithms. The following sections summarize the design choices and the final configuration used for the reported results.

3.3.1 Audio Features Extraction

We use YAMNet (Hershey et al., 2017) to produce 521-class scores and retain 16 classes selected for their relevance to crowd excitement and event cues. The selected classes were: 'Speech', 'Whoop', 'Singing', 'Choir', 'Chant', 'Whistling', 'Clapping', 'Applause', 'Chatter', 'Crowd', 'Air horn, 'Truck horn', 'Ambulance (siren)', 'Siren', 'Foghorn', 'Whistle', 'Chorus effect'. Regarding commentaries, they spanned multiple languages and when needed we translate/transcribe them for downstream text features (Gautam et al., 2024). Scores are min-max normalized and downsampled (aggregated bins at ≈4.8 s) to form the audio feature time series. These are multivariate time series, with 16 feature values generated every 0.48 seconds that are then downsampled using a moving average window, producing new values at intervals of 4.32 seconds, with each aggregated time point representing approximately the 4.8 second window. The coarser representation is better suited than the shorter frame-to-frame resolution (YAMNET defautl = 0.96 s window) to capture events meaningful in the context of soccer broadcasts.

We apply this procedure to each half of the games in SoccerNet, generating a tabular dataset. We then add a binary label column (1: highlight, 0: non-highlight) using the anchor timestamps provided in SoccerNet. To account for the temporal extension of highlights, we label not only the row corresponding to the SoccerNet anchor as a highlight, but also one adjacent row before and two after it. In this context, an "anchor" is the SoccerNet event timestamp mapped to the aggregated time series and a "row" is one aggregated time bin (≈4.8 s). Labeling the anchor plus one previous and two subsequent rows means identifying the aggregated bin containing the anchor (index 0) and marking indices -1, 0, +1 and +2 as positive (highlight). With the chosen temporal aggregation (4.32 s step, ≈4.8 s window) this labeling covers ≈17.8 s (≈±8.9 s) around the anchor, giving practical tolerance for annotation timing variability.

3.3.2 Low-Level Features

As a component of this study, we evaluated a small set of low-level acoustics (pitch, loudness, speech-rate) extracted with Praat/Librosa (Boersma and van Heuven, 2001; McFee et al., 2015).

We investigated the potential of these low-level acoustic features to complement and enhance highlight detection when combined with high-level features. As we show further on, we concluded that that was not the case. Moreover, the inclusion of these features in the final model would substantially increase the model's inference time, potentially limiting its real-time application capabilities.

3.3.3 Text Features

To incorporate information from the broadcast commentaries in the dataset , a NLP based feature is extracted and added to the features produced in the previous section. This feature extraction pipeline is divided into two steps: transcription and classification.

For the transcription step, the input audio is divided in blocks of 25.92 seconds and, for each block, OpenAI's Whisper_v3 automatic speech recognition model (Radford et al., 2023) is used to transcribe the broadcast speech in that block to text. By setting the language parameter to "english", the model will also translate the speech, generating all transcription text

in English.

In the classification step, the chunks of text from each block are classified using the zero-shot text classifier deberta-v3-large-zeroshot-v2.0 model (Laurer et al., 2024) as "highlight" or "non-highlight". The labels the model uses to classify these classes are "an action-packed or critical event, highlighting a moment that could significantly impact the match result" for "highlight" and "the match atmosphere, player substitutions, or comments on strategy without mentioning any key events" for "non-highlight", with both labels being prefixed by "This commentary from a soccer game describes" (hypothesis, reformulation from the natural language inference task for zero-shot classification).

The probability of the class "highlight" for each block (text score) will be used as the text feature for final highlight detection. Each block of 25.92 contains 6 blocks of 4.32 seconds (blocks from audio feature extraction), therefore, each score provided by the text feature extraction is duplicated for each corresponding audio block. The new dataset now has 16 audio features, 1 text feature and 1 label for each video segment of 4.8 seconds.

3.3.4 Supervised Combination of Audio and Text Features

Data are split by match into 500 games for training and 50 games held out for testing. For tuning we split the training set into train/validation (approx.82.4%/17.6%) stratified by class (seed `random_state=42`). We use scikit-learn's Gradient Boosting on the combined features. The trained model outputs a probability time series that is smoothed and processed with adaptive thresholding, peak detection and interval merging to produce final highlight segments. In more detail, post-processing procedure comprises the following steps: Signal smoothing, applying a rolling average with a window size of 7 and step size of 1 (the window corresponds to roughly 7×4.8 s ≈ 34 s of audio, which balances short transients and sustained crowd reactions); Simple adaptive thresholding, implementing an iterative approach to define a dynamic threshold with a scaling coefficient; Peak detection and interval identification, locating peaks in the smoothed signal and determining surrounding intervals according to amplitude and minimum-distance criteria.; Interval merging: Combining closely spaced intervals to form cohesive highlight segments.

3.4 Supervised Learning in Low-Dimensional Space

Supervised learning in this framework is only used to project the selected pretrained features into a lower-dimensional decision space using labeled examples. This design improves generalizability across leagues and can accommodate events that are not exhaustively labeled in SoccerNet but correlate with fan interest. Because the supervised component operates on compact, pretrained features, the framework can be adapted to other sports or esports with a small appropriate labeled dataset.

3.5 Evaluation Metrics

To assess the model's effectiveness, we report the standard metrics: precision, recall, and F1-score (Cady, 2017), together with two operational measures, coverage and bad coverage. Coverage is the fraction of the total broadcast time that is covered by the detected highlight intervals and helps quantify how concise the generated highlight reel is. Bad coverage is the fraction of the total clip time identified as highlights that, upon manual review, were deemed uninteresting to fans.

It is important to acknowledge that SoccerNet annotations are an imperfect proxy for what might interest an average fan: some events overlap (for example, a 'Shot on Target' can also be a 'Goal') and annotation timing can vary (subjective behaviour of the annotators). During error analysis we inspected representative mistakes and, where appropriate, adjusted interpretations of confusion metrics to account for clearly uninteresting labeled events. Specifically, we disregarded False Negative classifications for events that, upon manual review, were deemed uninteresting, even if they were labeled as highlights in the SoccerNet dataset. These manual checks reduce bias introduced by annotation noise, but do not replace the quantitative metrics reported on the held-out test set. The discrepancy arises from the imperfect nature of the SoccerNet labels and the subjective nature of what constitutes a "highlight" in soccer. Our manual review process allowed us to better align the model's output with human judgment of interesting events, suggesting that the model's real-world performance may be superior to what the standard metrics indicate.

4 EXPERIMENTS

The final model emerged after conducting a series of experiments in which several key factors influ-

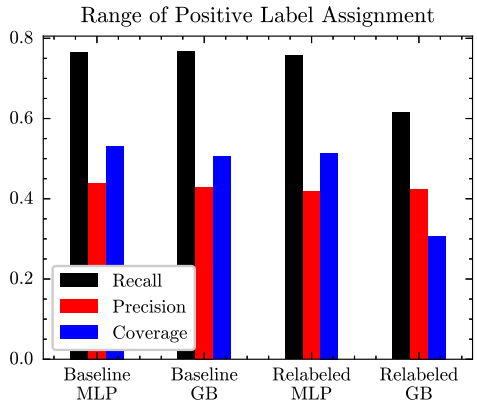

Figure 2: Effect of reducing the range of positive label assignment in the dataset, on the performance and coverage of the highlight detector, with Gradient Boosting (GB) and Multilayer Perceptron (MLP) as supervised classifiers. SoccerNet provides timestamps for each highlight, which we expand into temporal windows of different ranges.

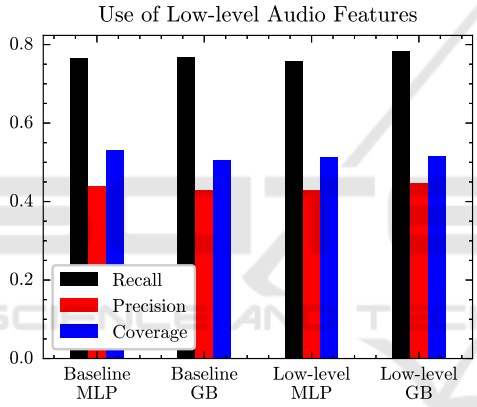

Figure 3: Impact of incorporating low-level audio features, in addition to selected YAMNET scores and a text feature, on the performance and coverage of the highlight detector, with Gradient Boosting (GB) and Multilayer Perceptron (MLP) as supervised classifiers.

enced its performance compared to that of the initial candidate. First, we compared different supervised model types, specifically a Multilayer Perceptron and a Gradient Boosting model. Second, we adjusted the range of positive label assignments around the SoccerNet anchor (1 row-anchor-2 rows versus 2-anchor-4 rows), as illustrated in Figure 2. Third, we analyze the incorporation of low-level audio features, as detailed in Figure 3. In addition, we explore various types of thresholds, as illustrated in Figure 4. Finally, we tuned a multiplicative coefficient applied to the threshold on validation splits.

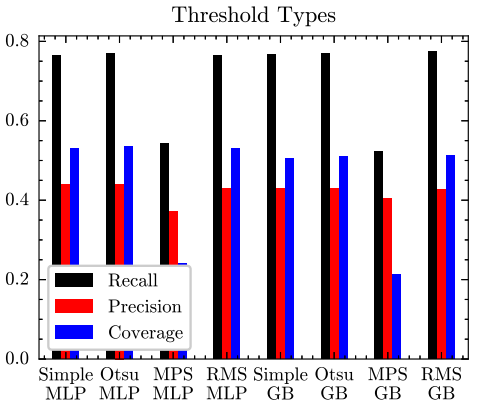

Figure 4: Effect of using different threshold functions in the peak detector on the performance and coverage of the highlight detector, with Gradient Boosting (GB) and Multilayer Perceptron (MLP) as supervised classifiers. Threshold functions considered include a simple adaptive threshold (Simple), Otsu's method (Otsu), mean plus one standard deviation (MPS) and root mean square (RMS).

5 RESULTS AND DISCUSSION

Table 1 reports the main results and summarizes our manual analysis performed on three full matches (see also Figure 5 for an example of first half match with SoccerNet anchors and predicted intervals). These qualitative checks revealed several important points. First, SoccerNet anchors do not always align with moments that human reviewers consider highlight-worthy, which introduces unavoidable label noise for this task. Second, relying on audio alone is challenging: crowd noise from routine team support can resemble excitement for eventful plays, and low-volume cheering (e.g., from a visiting-group) or subdued commentary can cause true highlights to be missed. Third, several model-predicted segments that were not labeled in SoccerNet corresponded to legitimately interesting near-miss events (e.g., shots off target), indicating that the model sometimes surfaces useful candidates that dataset labels omit.

Table 1: Benchmark model performance metrics.

Model Type	F1-Score	Precision	Recall	Coverage
Audio [1]	39.3%	33.0%	48.8%	20.3%
Audio + NLP	46.8%	44.0%	49.9%	19.0%

[1]For this model type a different threshold scaling coefficient (1.14) was used to obtain a comparable coverage. Coverage is the fraction of broadcast time covered by detected highlights, higher coverage indicates longer generated reels.

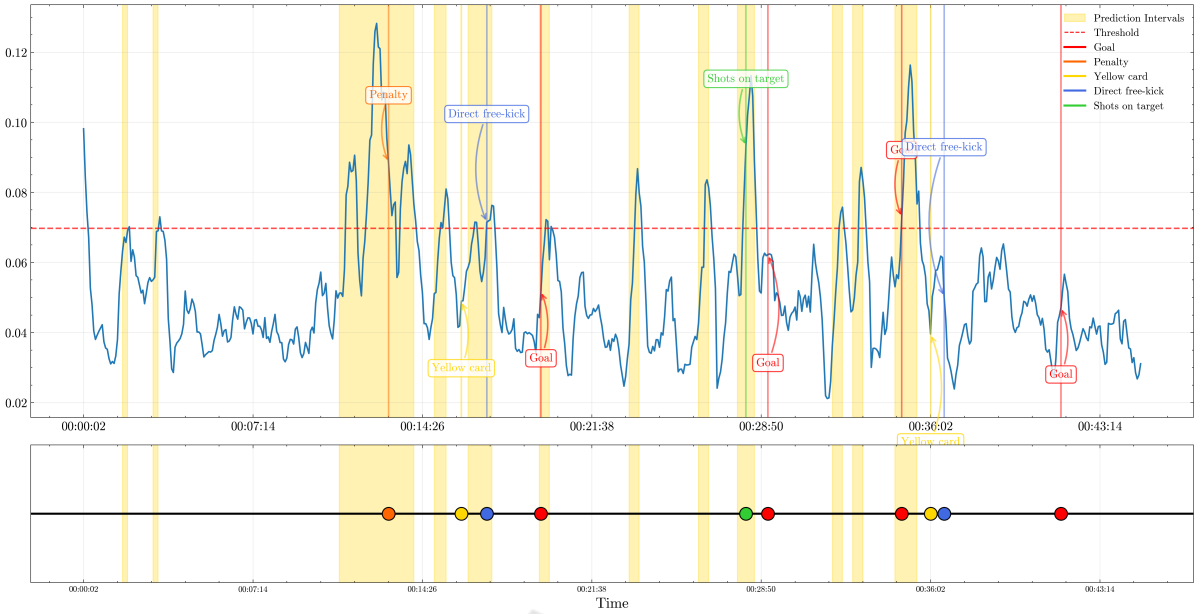

Figure 5: 2015-08-23 West Brom - Chelsea First-Half.

Table 2: Error analysis metrics.

Model Type	Precision	Recall	F1-score	Coverage	Bad coverage
Audio + NLP	47.6%	63.8%	54.5%	19.4%	28.1%

5.1 Discussion of Results

The following points summarize the main observations from our quantitative and qualitative analyses:

Labeling and features: Figure 2 shows that concentrating positive labels around SoccerNet anchors (the relabeling strategy described in Section 3) reduces coverage and improves class separation on validation data. Our controlled comparisons (Figure 3) found that adding the low-level acoustic features we extracted (notably `f0_mean`, `rate_of_speech` and `loudness_mean`, plus spectral bands `hf500` and `hf1000`) did not yield consistent improvements in the reported metrics while increasing inference cost. In development we observed that `f0_mean` and `loudness_mean` correlated with crowd reaction peaks, but their inclusion did not improve the end-to-end F1 on held-out data. This finding contrasts with previous studies that relied heavily on such features (Harb and Chen, 2003) (Baijal et al., 2015) (Xiong et al., 2003) (Raventós et al., 2015) (Dagtas and Abdel-Mottaleb, 2001) (Duxans et al., 2009), suggesting the superiority of our high-level feature approach for the context of this work.

Model selection and thresholding: We used simple adaptive thresholding for peak detection and tuned a single scaling coefficient on validation splits. The chosen value (1.17) trades off lower coverage for improved recall/precision on validation data. Across validation splits the Gradient Boosting classifier consistently outperformed the MLP variants we tested (Figures 2, 3 and 4), which motivated the final choice described in Section 3.3.4.

Performance and interpretation: Table 1 shows that adding the NLP-derived text feature improves F1 (46.8% vs 39.3%) and precision (44.0% vs 33.0%) compared to the audio-only baseline while producing a comparable coverage. These gains indicate the value of complementary signals from commentary transcriptions. The manual error analysis summarized in Table 2 further suggests that some model false positives (relative to SoccerNet labels) are semantically meaningful highlights under human inspection, suggesting our model captures relevant aspects beyond the original annotations. However, these manual checks are limited in scale and are reported as qualitative evidence rather than a replacement for the quantitative metrics on the held-out test set.

Together, these results support a practical lightweight multimodal pipeline that leverages pretrained deep models (YAMNet for high-level audio event scores and Whisper/NLI-based classification for commentary) combined with a compact supervised classifier. The use of pretrained deep feature extractors provides a richer semantic representation than classical handcrafted features, while the supervised classifier keeps inference efficient.

5.2 Limitations

Audio-only signals from TV broadcasts provide a useful but incomplete view. They are sensitive to league-dependent crowd behavior, commentator expressiveness, microphone placement, mixing levels, and broadcast production practices. Cultural differences in how fans and commentators express excitement can affect both audio and text signals. These factors limit out-of-domain generalization and motivate a multimodal approach (audio + video and/or game logs) for production-grade highlight systems.

Our manual error analysis was small-scale and intended to illustrate types of annotation noise and useful false positives, it does not substitute for broader re-annotation or additional benchmark datasets. While we tuned a single threshold-scaling parameter and report validation-based choices, full deployment should include per-broadcaster calibration and runtime measurements on representative hardware — this is left as future engineering work.

The architecture in this paper was designed to be adaptable and broadly applicable beyond soccer, including to eSports. Many competitive titles exhibit commentary and event-driven audio patterns similar to broadcast sports—shoutcasters, structured commentary, and discrete in-game events—so the proposed multi-modal approach can be relevant across domains. Nonetheless, applying the methodology to a specific eSports title requires careful validation because domain shifts (synthesized in-game sounds, frequent UI/notification cues, different shoutcaster conventions, or absent crowd noise) can materially change pretrained-model feature responses. We therefore recommend treating eSports transfer as a domain-adaptation task: collect representative labelled samples for each target title, validate feature suitability (and recalibrate or fine-tune pretrained components where necessary), and run held-out evaluations to measure coverage and false-positive modes specific to the game. A pilot study with labelled matches and per-title validation is advised before production deployment for automated highlight generation in eSports.

6 CONCLUSION

Audio signals complement video for highlight detection and, when combined with commentary-derived text, improve accuracy over audio-only baselines. We find that high-level pretrained audio scores (YAMNet) plus a lightweight supervised classifier yield a compact, interpretable feature set that suits low-latency deployments and often outperforms handcrafted low-level features.

Future research could explore ways to adapt the audio-based classification model to account for the contextual variables referred in the precedent section, potentially through transfer learning techniques or the incorporation of meta-data about the league, broadcaster, or cultural context. One immediate approach to enhancing the current model's performance could be to relabel the dataset using fewer, but more appropriate classes from SoccerNet that better serve as proxies for highlights. A more labor-intensive option would involve manually labeling a smaller, curated dataset with labels tailored to the specific needs of the application, relying on user-based judgment rather than solely depending on the predefined SoccerNet labels.

Although validated on SoccerNet, the multimodal approach is portable to other broadcast domains (including eSports) provided representative target data are collected for calibration and validation.

ACKNOWLEDGEMENTS

This work was supported under the base funding project of the DTx CoLAB - Collaborative Laboratory, under the Missão Interface of the Recovery and Resilience Plan (PRR), integrated in the notice 01/C05-i02/2022, which aims to deepen the effort to expand and consolidate the network of interface institutions between the academic, scientific and technological system and the Portuguese business fabric. We thank our Portuguese industry partners for their collaboration. To protect intellectual property, some experiments were performed on confidential datasets and are not reported here. The experiments and results presented in this paper were produced using publicly available data to ensure reproducibility.

REFERENCES

Baijal, A., Cho, J., Lee, W., and Ko, B.-S. (2015). Sports highlights generation based on acoustic events detection: A rugby case study. In *2015 IEEE International Conference on Consumer Electronics (ICCE)*, pages 20–23.

Boersma, P. and van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, 5(9).

Boes, W. and Van hamme, H. (2021). Audiovisual transfer learning for audio tagging and sound event detection. In *Interspeech 2021*.

Cady, F. N. (2017). *The Data Science Handbook*. John Wiley and Sons.

Dagtas, S. and Abdel-Mottaleb, M. (2001). Extraction of TV highlights using multimedia features. In *2001 IEEE Fourth Workshop on Multimedia Signal Processing (Cat. No.01TH8564)*, pages 91–96.

Deliege, A., Cioppa, A., Giancola, S., Seikavandi, M. J., Dueholm, J. V., Nasrollahi, K., Ghanem, B., Moeslund, T. B., and Van Droogenbroeck, M. (2021). SoccerNet-v2: A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4503–4514.

Duxans, H., Anguera, X., and Conejero, D. (2009). Audio based soccer game summarization. In *2009 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pages 1–6.

Filippidis, P.-M., Dimoulas, C., Bratsas, C., and Veglis, A. (2018). A Multimodal Semantic Model for Event Identification on Sports Media Content. *Journal of Media Critiques*, 4(14):295–306.

Gautam, S., Sarkhoosh, M. H., Held, J., Midoglu, C., Cioppa, A., Giancola, S., Thambawita, V., Riegler, M. A., Halvorsen, P., and Shah, M. (2024). SoccerNet-Echoes: A Soccer Game Audio Commentary Dataset.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Giancola, S., Amine, M., Dghaily, T., and Ghanem, B. (2018). SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1792–179210.

Harb, H. and Chen, L. (2003). Highlights Detection in Sports Videos Based on Audio Analysis. In *Third International Workshop on Content-Based Multimedia Indexing, CBMI03*, pages 1–7, Rennes, France.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. (2017). Cnn architectures for large-scale audio classification.

Laurer, M., van Atteveldt, W., Casas, A., and Welbers, K. (2024). Building Efficient Universal Classifiers with Natural Language Inference.

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., and Nieto, O. (2015). Librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*.

Midoglu, C., Sabet, S. S., Sarkhoosh, M. H., Majidi, M., Gautam, S., Solberg, H. M., Kupka, T., and Halvorsen, P. (2024). AI-Based Sports Highlight Generation for Social Media. In *Proceedings of the 3rd Mile-High Video Conference on Zzz*, pages 7–13, Denver CO USA. ACM.

Nergård Rongved, O. A., Stige, M., Hicks, S. A., Thambawita, V. L., Midoglu, C., Zouganeli, E., Johansen, D., Riegler, M. A., and Halvorsen, P. (2021). Automated Event Detection and Classification in Soccer: The Potential of Using Multiple Modalities. *Machine Learning and Knowledge Extraction*, 3(4):1030–1054.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 28492–28518, Honolulu, Hawaii, USA. JMLR.org.

Raventós, A., Quijada, R., Torres, L., and Tarrés, F. (2015). Automatic summarization of soccer highlights using audio-visual descriptors. *SpringerPlus*, 4(1):301.

Ringer, C., Nicolaou, M. A., and Walker, J. A. (2022). Autohighlight: Highlight detection in league of legends esports broadcasts via crowd-sourced data. *Machine Learning with Applications*, 9:100338.

Rui, Y., Gupta, A., and Acero, A. (2000). Automatically extracting highlights for TV Baseball programs. In *Proceedings of the Eighth ACM International Conference on Multimedia*, MULTIMEDIA '00, pages 105–115, New York, NY, USA. Association for Computing Machinery.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition.

Valliappan, N. H., Pande, S. D., and Reddy Vinta, S. (2024). Enhancing gun detection with transfer learning and yamnet audio classification. *IEEE Access*.

Vanderplaetse, B. and Dupont, S. (2020). Improved Soccer Action Spotting using both Audio and Video Streams. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3921–3931, Seattle, WA, USA. IEEE.

Xiong, Z., Radhakrishnan, R., Divakaran, A., and Huang, T. (2003). Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 5, pages V–632.

Xiong, Z. and Wang, H. (2023). Dual-stream multimodal learning for topic-adaptive video highlight detection. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*.