Illegal waste dumping detection

Thierry Bouwmans
Laboratoire MIA
La Rochelle Université, France
thierry.bouwmans@univ-lr.fr

Antonio Greco
DIEM Department
University of Salerno, Italy
agreco@unisa.it

Sébastien Piérard
Montefiore Institute
University of Liège, Belgium
s.pierard@uliege.be

Andrea Vincenzo Ricciardi
DIEM Department
University of Salerno, Italy
anricciardi@unisa.it

Carlo Sansone
DIETI Department
University of Napoli Federico II, Italy
carlo.sansone@unina.it

Marc Van Droogenbroeck
Montefiore Institute
University of Liège, Belgium
m.vandroogenbroeck@uliege.be

Bruno Vento
DIETI Department
University of Napoli Federico II, Italy
bruno.vento@unina.it

Abstract

Illegal waste dumping represents a serious environmental and public health challenge, motivating the development of automated surveillance systems capable of detecting such events in real time. This paper describes and analyzes the results of the Illegal Waste Dumping Detection (IWDD) international contest, which aims to advance video-based methods for recognizing illegal disposal activities from fixed surveillance cameras. We describe the Mivia-IWDD dataset introduced for the competition, consisting of 400 video clips (200 positive, 200 negative) with precise temporal annotations, covering both static and dynamic dumping actions as well as challenging negative scenarios across diverse environmental conditions. Ten teams participated in the contest, proposing heterogeneous approaches based on spatio-temporal deep learning, action recognition, temporal modeling, and efficiency-oriented design choices. We evaluated the methods using a comprehensive protocol that combines classical detection metrics (Precision, Recall, and F_1-score) with additional indicators targeting real-time applicability, including notification delay, processing frame rate, and memory usage. Moreover, we analyzed and compared the results achieved by all teams from multiple perspectives. Beyond ranking performance, this paper provides useful insights and highlights open challenges and promising research directions, contributing a benchmark and practical guidelines for future work on illegal waste dumping detection in smart surveillance systems.

1. Introduction

Illegal waste dumping is a critical threat to communities worldwide. Beyond aesthetic concerns, it strains municipal budgets, damages ecosystems, and endangers public health [24, 27], ultimately undermining urban sustainability and residents' quality of life [3, 18, 26]. A major barrier to effective deterrence is the lack of sophisticated monitoring technologies. Current research is constrained by a "data void": existing datasets mainly consist of static images capturing waste after disposal, failing to represent the crucial dynamic moments of the act itself. Without video data reflecting the diversity of human behaviors, waste types, and environmental conditions, the development of proactive real-time intervention systems remains limited.

To address this critical gap, we organized the Illegal Waste Dumping Detection (IWDD) contest, promoting the development of automated video-based surveillance systems capable of detecting illegal waste disposal in real time from static videosurveillance cameras. Organizing competitions has recently demonstrated to significantly contribute to the state of the art progress in specific research topics [6–9]. Central to this initiative was the release of the MIVIA-IWDD dataset, consisting of 400 videos (200 positive, 200 negative) used as training set for the competition. This dataset captures the full temporal evolution of disposal events, encompassing both static and dynamic disposal modalities, and representing the diverse patterns and complexities of the real world. To ensure robustness, the collection includes challenging negative scenarios, which are common activities that resemble dumping behaviors. Provided with precise temporal annotations, the dataset offers a rigorous framework for training and validating methods to detect the exact onset of illegal disposal acts.

The ten methods submitted to the competition were evaluated using standard event-level detection metrics, including Precision, Recall, and F_1-score, complemented by real-time indicators such as notification delay, processing frame rate, and memory usage. The analysis of the performance has been done on a private test set, characterized by the same a priori distribution and illegal waste dumping actions, but collected in different scenarios. In this way, we evaluated the performance of the methods in a realistic framework, in which the generalization capability of the approaches is tested under real-world conditions.

Therefore, we can summarize the contributions of this work as follows. We introduce MIVIA-IWDD, the first open-source evenly balanced training set capturing static and dynamic illegal waste dumping actions under diverse conditions. We describe the IWDD contest, specifically organized to benchmark illegal waste dumping detection methods trained with this dataset. We report the results obtained by the submitted methods evaluating them on a private test set collected in real-world scenarios, proposing an evaluation protocol combining classical metrics with real-time performance indicators and analyzing the results from multiple perspectives. By analyzing the submitted approaches, we highlight useful insights, offering a benchmark and guidance for future research activities in illegal waste dumping detection.

2. Related works

The scarcity of publicly available benchmarks for the IWDD task has forced researchers to rely mostly on private custom-built collections. Existing resources can be broadly categorized by data modality: image-based datasets, used primarily for waste detection, and video-based datasets, used for action recognition. Table 1 provides a comparative overview of these resources.

2.1. Image datasets

Image datasets generally support supervised learning tasks aimed at recognizing waste objects in static scenes.

TACO [22] is a prominent open-source dataset containing thousands of images with high-quality segmentation masks, covering diverse environments from beaches to streets. Similarly, pLitter [16] focuses on plastic waste, offering specific subsets for roadside and floating river debris. In the broader context of environmental monitoring, remote sensing resources such as AerialWaste [25] and the dataset by Marrocco et al. [17] utilize aerial and satellite imagery. However, these are designed to identify accumulated illegal dumpsites rather than real-time dumping events.

Despite their utility for waste inventory and cleanup, static imagery is inherently insufficient for the specific requirements of illegal waste dumping detection. The core challenge of this task lies not in detecting the waste, but in recognizing the action (abandonment). A single frame cannot disambiguate between a pre-existing pile of trash, an object currently being carried, or an item being discarded. Temporal dynamics are essential, as distinguishing illegal dumping from mere litter requires analyzing motion and context over time.

2.2. Video datasets

Video datasets provide the necessary temporal context to model and recognize the action of dumping, distinguishing it from the static presence of waste. As summarized in Table 1, existing works have addressed this problem by collecting footage focused on specific dumping modalities.

Mahankali et al. [15] focused exclusively on dynamic disposal with a dataset of 50 videos depicting waste discarded from moving vehicles. Their collection is strictly limited to vehicle-based infractions, explicitly excluding instances where individuals dump waste on foot. In the context of surveillance, Yun et al. [29] utilized CCTV footage from eight locations to capture changes in the spatial relationship between pedestrians and objects. Similarly, Husni et al. [11] developed a larger dataset of 400 videos in Indonesia (market and river areas), featuring a strictly balanced distribution of 200 littering and 200 non-littering events. Other contributions target the fine-grained mechanics of disposal to support pose-based analysis. Kim et al. [14] created 30 short clips simulating eight specific postures (e.g., one-handed, two-handed, without bending) to analyze wrist and body joints. Similarly, Yu et al. [28] compiled 72 videos depicting diverse dumping intensities, ranging from carefully lowering waste to forcefully throwing it.

Despite the methodological diversity of these contributions, a critical shortcoming shared by all existing video datasets is their unavailability to the public. They remain private assets, which prevents the research community from reproducing results and establishing a standardized benchmark. To address this gap, we organize the competition around the MIVIA-IWDD dataset, the first publicly released training set for illegal waste dumping detection.

3. Dataset

In this section, we present MIVIA-IWDD, the novel training set specifically designed to support the development of IWDD systems for the competition. We detail the data collection methodology, the annotation protocol, and the statistical properties that characterize the training set given to the teams. Then, we describe the private test set adopted to evaluate the performance of the approaches submitted by the contest teams.

Table 1. Overview of existing waste management datasets, showing data type (image or video), disposal modality, samples, and access.

	Dataset	Disposal modality	Samples	Access
Images	TACO (2020) [22]	Abandoned Waste	1,500 images	✓
	pLitter (2024) [16]	Abandoned Plastic Litters	12,752 images	✓
	AerialWaste (2023) [25]	Aerial-Abandoned Waste	11,703 images	✓
	Marrocco et al. (2024) [17]	Aerial-Abandoned Waste	4,715 images	✗
Videos	Mahankali et al. (2018) [15]	Dynamic Disposal	50 videos	✗
	Yun et al. (2019) [29]	Dynamic and Static Disposal	N/A	✗
	Husni et al. (2021) [11]	Dynamic and Static Disposal	400 videos	✗
	Kim et al. (2022) [14]	Static Disposal	30 videos	✗
	Yu et al. (2024) [28]	Dynamic and Static Disposal	72 videos	✗
	MIVIA-IWDD (2026)	**Dynamic and Static Disposal**	**400 videos**	✓

3.1. MIVIA-IWDD training set

3.1.1. Data collection

The primary design goal of MIVIA-IWDD was to create a resource that mirrors the complexity of real-world surveillance environments. Unlike synthetic datasets or those employing actors, our collection process prioritized ecological validity. Most of the footage was obtained from real surveillance cameras, capturing spontaneous and unscripted human behavior across a wide variety of scenes. Figure 1 offers a representative visual overview of the dataset.

The dataset is built around two core categories:

- **Positive samples**: We ensured coverage of the full spectrum of disposal behaviors. This includes *static disposal*, where individuals deliberately place objects such as furniture, bags, or construction debris on the ground, and *dynamic disposal*, which involves throwing waste from moving vehicles or while walking. These scenarios present distinct challenges: static events often occur in cluttered environments with potential occlusions, while dynamic events require the system to detect rapid motion and temporal anomalies.
- **Negative samples**: To train a robust discriminator, it is insufficient to simply include empty scenes. Therefore, we meticulously selected hard negative videos depicting actions that are visually similar to dumping but entirely legitimate. These include ambiguous behaviors (e.g., carrying bulky items without discarding them), scenarios where pedestrians pass near pre-existing waste, and authorized sanitation operations. This variety ensures may help the model to distinguish actual violations from lawful waste management or unrelated daily activities.

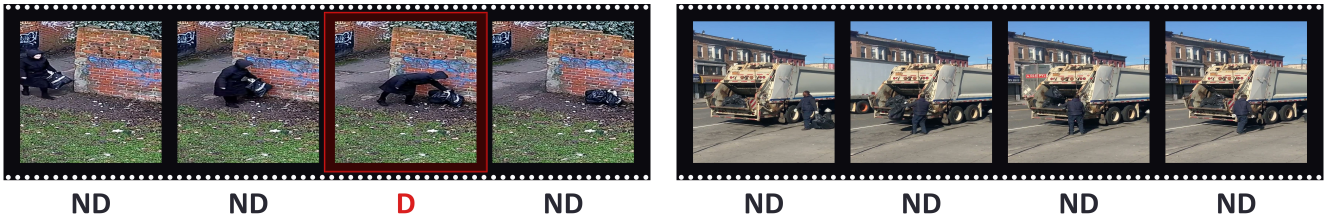
3.1.2. Dataset annotation

To ensure high-quality ground truth, the dataset underwent a rigorous manual annotation process. Human annotators reviewed each video to verify the class label (Dumping vs. No Dumping). Positive samples include a precise temporal annotation marking the event onset, enabling reaction-time evaluation and action spotting [5], which localizes actions in untrimmed videos using a single timestamp. Each event is also categorized as static or dynamic. Additionally, metadata concerning environmental conditions, specifically the time of day (day/night) and lighting quality (bright/dim), were tagged for every clip. This rich metadata layer facilitates in-depth error analysis and allows researchers to assess model performance across different domains.

3.1.3. Dataset statistics

The MIVIA-IWDD dataset comprises a total of 400 videos, amounting to over 2 hours of footage. In terms of technical specifications, the recordings predominantly feature high definition resolutions (1920 × 1080 and 1280 × 720) and frame rates between 20 and 30 FPS, aligning with standard modern surveillance configurations. A range of lower and higher resolutions is also included to simulate legacy and diverse hardware setups. The clips are concise, with an average duration of approximately 20 seconds, ensuring the content remains focused on the relevant window of activity.

A key strength of the dataset is its strict balance, meticulously designed to prevent algorithmic bias. The 400 videos are evenly split into 200 positive and 200 negative samples. The positive subset is further stratified to include exactly 100 static and 100 dynamic disposal events, ensuring equal representation of both modalities. Regarding environmental diversity, the dataset includes a realistic proportion of



(a) Illegal waste dumping.

(b) No dumping.

Figure 1. Sample video frames from the MIVIA-IWDD dataset illustrating different scenarios. The examples include both static and dynamic disposal events, alongside various negative videos depicting non-dumping activities. Frames marked with D indicate the key moments in which the dumping action occurs and becomes evident, while frames labeled ND represent non-dumping situations.

lighting conditions, with a subset of videos captured during nighttime or low-light scenarios, which are critical for 24/7 surveillance applications.

3.2. Private test set

The test set is kept confidential and is not accessible to participants, guaranteeing an unbiased and realistic assessment of the generalization ability of the proposed methods. It consists of 100 video samples, including 50 positive cases (25 with static disposal and 25 with dynamic disposal) and 50 negative cases, preserving the same prior class distribution as the training data. The test scenarios differ from those used for training, allowing the evaluation of the approaches under real-world conditions that are not represented in the training set. The final ranking of the contest is determined by the overall performance achieved on the entire test set.

4. Methods

This section summarises the methods submitted to the contest for illegal waste dumping detection. Although all approaches address the same task under a shared evaluation protocol, they adopted markedly different architectural designs and temporal reasoning strategies, ranging from deep learning models with explicit sequence modelling to multimodal vision-language pipelines. These methodological choices reflect distinct trade-offs between *localisation accuracy*, *robustness*, and *computational efficiency*. We report the methods in alphabetical order of the team name.

The **AGH-EVS** approach formulates illegal dumping detection as a *frame-level temporal action detection* problem using a causal spatio-temporal convolutional network. Built on a pre-trained MoViNet backbone operating in causal mode, the method produces frame-wise dumping probabilities through a lightweight temporal prediction head. Inference relies on a sliding-window strategy, where short clips are analysed independently and their predictions aggregated to form a dense temporal confidence signal over the entire video. The dumping timestamp is determined by the frame with the highest confidence, subject to a fixed detection threshold. This fully convolutional formulation enables effective temporal localisation without resorting to recurrent models or explicit proposal generation.

The **ALEXKNIGHTS** [1] approach addresses illegal dumping detection as a *binary video classification task* using an efficient spatio-temporal convolutional architecture. The method builds on the lightweight X3D-M backbone, which employs depthwise separable 3D convolutions and channel-wise attention to capture discriminative spatio-temporal patterns at low computational cost. The pre-trained multi-class classification head is replaced with a binary classifier, and fine-tuning is performed using differential learning rates to balance task adaptation and feature preservation. During training, clips are sampled around annotated dumping events for positive examples and from temporally distant regions for negatives to increase data diversity. At inference time, temporal localisation is achieved through a sliding-window strategy, selecting the most confident window and returning its centre timestamp when the confidence exceeds a predefined threshold.

The **ARRAY** [12] approach formulates dumping detection as a frame-level temporal classification problem grounded in *multimodal semantic representations*. Videos are temporally subsampled, and each frame is independently encoded through a vision–language pipeline that generates multiple descriptive captions capturing scene context, human actions, and waste-related cues. These captions are embedded using Sentence-BERT to obtain compact semantic representations, which are then analysed by a temporal classifier. During inference, frame-level probabilities are temporally smoothed, and dumping events are identified using confidence and duration based thresholds. Training employs a weighted binary cross-entropy loss to address class imbalance, with temporal subsampling acting as the primary preprocessing step.

The **GARBERUS** [23] method adopts a *multi-task learning framework* built upon a shared spatio-temporal backbone to jointly address illegal dumping detection, temporal localisation, and disposal-type classification. The approach leverages a pre-trained TimeSformer architecture to extract spatio-temporal representations, which are processed by three independent binary heads following a divide-and-conquer strategy: a primary dump versus no-dump classifier, a temporal localisation head providing coarse timestamp estimates, and a disposal-type classifier distinguishing between static and dynamic dumping events. Task-dependent temporal sampling is used to balance efficiency and temporal coverage. Robustness is further improved through additional training samples for the primary task, data augmentation, and early stopping, enabling the unified handling of multiple dumping-related objectives within a single architecture.

The **IMSLAB** [2] method tackles illegal waste dumping detection and onset localisation through a *multi-stage onset-centric temporal reasoning framework* designed for untrimmed surveillance videos. Frame-level visual representations are extracted from short uniformly sampled clips using a frozen VideoMAE backbone, explicitly favouring boundary-sensitive modelling over long-term temporal pooling. Motion-aware descriptors are obtained by enriching appearance features with temporal differences and multi-scale temporal pooling, and are processed by a lightweight temporal module combining dilated temporal convolutions and self-attention. In addition to motion information, the method explicitly models *scene-state transitions* to capture persistent environmental changes indicative of dumping events. These complementary cues are fused into an onset evidence curve, which is exploited by dedicated classification and regression heads to separate genuine dumping actions from background motion and to estimate a normalised onset interval. At the end, predictions from multiple temporal views are combined using confidence-weighted voting, yielding robust video-level decisions and onset estimates.

The **SAFE:CAUTION** [13] method frames illegal dumping detection as a *video-level multi-task problem*, jointly addressing event classification and timestamp estimation. The approach relies on two complementary pre-trained video backbones, Hiera and VideoMAE V2, which are fine-tuned to extract global spatio-temporal representations from uniformly sampled clips. For each backbone, lightweight classification and regression heads are employed to predict a dumping probability and a normalised event timestamp, respectively. Training is driven by a multi-task objective that combines binary cross-entropy with mean squared error, with the regression loss applied only to positive samples. Final predictions are made via weighted ensembling and thresholding, with timestamps mapped to the original video duration.

The **SUDOPI** [30] method follows a *CV pipeline* that integrates object detection, multi-object tracking, background modelling, and rule-based temporal reasoning. A custom-trained YOLOv8 detector is configured to prioritise Recall, enabling the detection of small or partially occluded waste objects, which are subsequently associated over time using the BoT-SORT tracker. To distinguish newly deposited waste from static scene elements, the approach employs a dual-timescale MOG2 background model that captures both short-term changes and long-term scene stability. Candidate dumping events are validated using multiple novelty cues and strict temporal consistency rules enforcing object persistence and stationarity, while additional exclusion heuristics suppress background clutter. This sequential detection, validation, and temporal filtering pipeline yields the final dumping decision and corresponding timestamp.

The **UNICASS** [4] method adopts a two-stage deep learning architecture that integrates frame-level spatial feature extraction with recurrent temporal modelling. Visual features are extracted using a ResNet-18 backbone pretrained on ImageNet, yielding fixed-dimensional embeddings for each frame, which are subsequently processed by a bidirectional GRU to capture temporal context. Frame-level dumping probabilities are then produced by a lightweight classification head. To encourage effective temporal learning, several frame-level supervision strategies are explored during training, while inference strictly adheres to the challenge annotation protocol. Given the short duration of the videos, inputs are represented as fixed-length sequences of 64 frames, uniformly sampled and padded when required to ensure temporal consistency and computational efficiency.

The **WASTEBUSTERS** method employs a *two-stage pipeline* that combines frozen pre-trained video representations with lightweight temporal sequence modelling. Videos are partitioned into temporally ordered clips processed at low frame rate and encoded using a pre-trained VideoMAEv2 transformer. The resulting clip-level embeddings are analysed by a compact Mamba state-space sequence model, which propagates temporal context across clips to produce dumping predictions. During inference, clip-level outputs are temporally smoothed to reduce noise before generating final video-level decisions, resulting in an efficient and robust detection pipeline.

The **WASTESCRAPPER** method combines *frozen vision–language representations* with a lightweight temporal localisation network to perform dumping detection and timestamp estimation. Videos are uniformly sampled into a fixed number of frames and encoded using a pre-trained CLIP image encoder. The resulting frame-level embeddings are processed by a compact temporal model composed of one-dimensional convolutions and self-attention layers to capture temporal dependencies. A video-level classification score is obtained via temporal pooling, while the dumping timestamp is estimated directly from the temporal score distribution. At inference time, predictions are thresholded and timestamps are mapped back to the original video duration.

5. Evaluation metrics

According to the contest rules, the submitted methods are evaluated under a unified protocol designed to assess both detection accuracy and suitability for real-time deployment.

5.1. Event-based detection metrics

Detection performance is evaluated by comparing the predicted dumping instant p with the ground-truth g, defined as the first frame in which the dumping action becomes visible. Early detections are tolerated within a temporal margin of $\Delta t = 5$ seconds, while detections occurring more than $T_{\max} = 10$ seconds after g are considered invalid.

A detection is counted as a **true positive** if it occurs in a positive video and falls within the valid temporal window $(g - \Delta t) \leq p \leq (g + T_{\max})$. Detections produced in negative videos, as well as detections outside this interval in positive videos, are treated as **false positives**. Conversely, positive videos for which no valid detection is generated are counted as **false negatives**.

Based on these definitions and being TP, FP and FN the number of true positives, false positives and false negatives, standard event-level metrics are computed, namely Precision, Recall, and F_1 score:

$$\begin{aligned}
\mathrm{P} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \\
\mathrm{R} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \qquad (1)\\
\mathrm{F}_1 &= 2 \times \frac{\mathrm{P}\times\mathrm{R}}{\mathrm{P}+\mathrm{R}}
\end{aligned}$$

The final ranking is determined by the highest F_1 score on the test set, which provides a balanced value by jointly accounting for false alarms and missed dumping events. While the F_1 score provides an effective trade-off between precision P and recall R, it does not fully capture the relative importance or ranking implied by different combinations of these two metrics [21]. This limitation motivates the need to evaluate models from an alternative perspective, as discussed in Section 6.2.

5.2. Efficiency and real-time metrics

Beyond detection accuracy, the evaluation protocol includes metrics designed to assess the *practical suitability* of the proposed methods for real-time surveillance scenarios, particularly in settings with limited computational resources, with an emphasis on detection latency and computational efficiency.

Notification delay. This metric evaluates how promptly a method identifies an illegal dumping event after its onset. For each correctly detected positive video, the delay is defined as $d_i = |p_i - g_i|$. The average delay D computed over all true positives is normalised as

$$\mathrm{D}_{\text{norm}} = \max\left(0, 1 - \frac{\mathrm{D}}{\mathrm{T}_{\max}}\right), \qquad (2)$$

where higher values correspond to faster and more timely detections.

Processing frame rate. Computational throughput is measured in terms of the average number of frames processed per second on a target GPU. Let PFR denote the measured processing frame rate and $\mathrm{PFR}_{\text{target}}$ a predefined reference value. The corresponding normalised score is defined as

$$\mathrm{PFR}_{\Delta} = \max\left(0, \frac{\mathrm{PFR}_{\text{target}}}{\mathrm{PFR}} - 1\right), \qquad (3)$$

with higher scores indicating more efficient processing.

Memory usage. Memory efficiency is assessed based on the peak GPU memory consumption observed during inference. Given a target memory budget $\mathrm{MEM}_{\text{target}}$ and the measured peak usage MEM, the normalised memory score is computed as

$$\mathrm{MEM}_{\Delta} = \max\left(0, \frac{\mathrm{MEM}}{\mathrm{MEM}_{\text{target}}} - 1\right), \qquad (4)$$

where lower memory usage results in better scores.

6. Results

6.1. Analysis with the contest metrics

The results of the competition, summarized in Table 2, highlight a significant variety of strategies and a clear trade-off between detection quality and computational efficiency among the participating teams. AGH-EVS achieves the highest F_1 score of 0.57, largely driven by a strong recall of 0.74, despite moderate precision at 0.47, suggesting that its frame-level temporal action detection effectively captures most dumping events but at the cost of false positives. GARBERUS balances precision and recall (both 0.54), yielding a solid F_1 of 0.54 and excellent normalised detection efficiency ($\mathrm{D}_{norm} = 0.8$) indicating that its multi-task TimeSformer architecture effectively leverages temporal context while maintaining computational efficiency. SAFE:CAUTION exhibits the highest precision on the test set (0.62) but suffers from low recall (0.32), resulting in an F_1 of 0.42; this indicates that its video-level multi-task ensemble, while highly selective, misses a significant fraction of true dumping events. ARRAY and IMSLAB show moderate performance (F_1 of 0.46 and

Table 2. Final competition ranking evaluated on the private test set, sorted by F$_1$-Score. Efficiency metrics are calculated using reference targets PFR$_{\text{target}} = 30$ and MEM$_{\text{target}} = 2$. Best results are in **bold**, second-best are underlined.

Team	Quality Metrics			Efficiency Metrics				
	P	R	F$_1$	D$_{\text{norm}}$ ↑	PFR ↑	PFR$_\Delta$ ↓	MEM ↓	MEM$_\Delta$ ↓
AGH-EVS	0.47	**0.74**	**0.57**	0.38	13.25	1.26	**0.12**	0.00
GARBERUS [23]	0.54	0.54	0.54	0.80	63.59	0.00	0.65	0.00
ARRAY [12]	0.49	0.44	0.46	0.74	0.81	35.92	1.08	0.00
IMSLAB [2]	0.47	0.38	0.42	0.79	268.73	0.00	0.50	0.00
SAFE:CAUTION [13]	**0.62**	0.32	0.42	**0.83**	4.53	5.61	2.88	0.44
WASTEBUSTERS	0.42	0.40	0.41	0.73	6.97	3.29	1.00	0.00
SUDOPI [30]	0.43	0.38	0.40	0.78	15.02	0.99	0.33	0.00
UNICASS [4]	0.39	0.30	0.34	0.67	**744.82**	0.00	0.53	0.00
WASTESCRAPPER	0.50	0.26	0.34	0.74	491.98	0.00	0.50	0.00
ALEXKNIGHTS [1]	0.50	0.06	0.11	0.73	25.73	0.16	0.15	0.00

0.42, respectively), reflecting the trade-offs inherent in multimodal frame-level encoding and onset-centric temporal reasoning. WASTEBUSTERS, SUDOPI, UNICASS, and WASTESCRAPPER all achieve similar F$_1$ scores ranging from 0.34 to 0.41, suggesting that recurrent temporal modelling, classical pipelines, and lightweight sequence models provide comparable detection coverage. ALEXKNIGHTS achieves negligible recall (0.06), causing its F$_1$ to collapse, highlighting the limitations of highly efficient video-level classification in capturing sparse events.

Table 3. Performance ranking evaluated on the training set, sorted by F_1-Score. The table includes also the confusion matrix metrics. Best results are in **bold**, second-best are underlined.

Team	Quality Metrics		
	P	R	F$_1$
SAFECAUTION [13]	0.95	**0.91**	**0.93**
ALEXKNIGHTS [1]	**0.97**	0.85	0.90
WASTESCRAPPER	0.89	0.83	0.86
WASTEBUSTERS	0.86	0.83	0.84
SUDOPI [30]	0.94	0.72	0.81
UNICASS [4]	0.87	0.76	0.81
GARBERUS [23]	0.79	0.77	0.78
ARRAY [12]	0.72	0.62	0.67
IMSLAB [2]	0.51	0.79	0.62
AGH	0.62	0.50	0.55

Comparing with the training set results (Table 3), it is evident that some methods such as SAFE:CAUTION and ALEXKNIGHTS exhibit strong overfitting, with F$_1$ scores above 0.90 on training data but significantly lower performance on the test set. In contrast, AGH-EVS and GARBERUS maintain a more consistent generalization gap, underlining the robustness of their temporal modelling strategies. ARRAY and IMSLAB exhibit moderate training set performance, in line with their test set results, implying that multimodal frame-level and onset-centric temporal reasoning approaches achieve stable but not outstanding generalization. WASTEBUSTERS, SUDOPI, UNICASS, and WASTESCRAPPER all show unbalanced training and test F$_1$ scores. These observations emphasize that while strong training set metrics indicate effective model capacity, final ranking is ultimately determined by test set generalization, underlining the importance of using a challenging and realistic test set.

6.2. Analysis from a different perspective

Considering that all the teams achieved a F$_1$ lower than 0.6 on the test set, with values of P and R lower than 0.5 in most of the cases, we analyzed the results also from a different perspective, namely by following the approaches proposed in [20] and [19]. We first examine the ROC space in Fig. 2, computed using the standard definition of false positives, as these approaches rely on the classical framework.

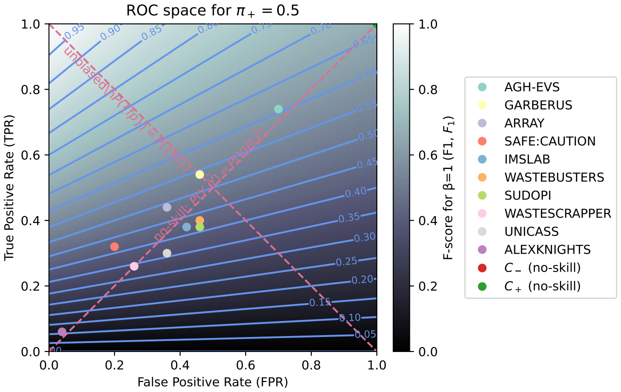


Figure 2. ROC space of the teams on the test set.

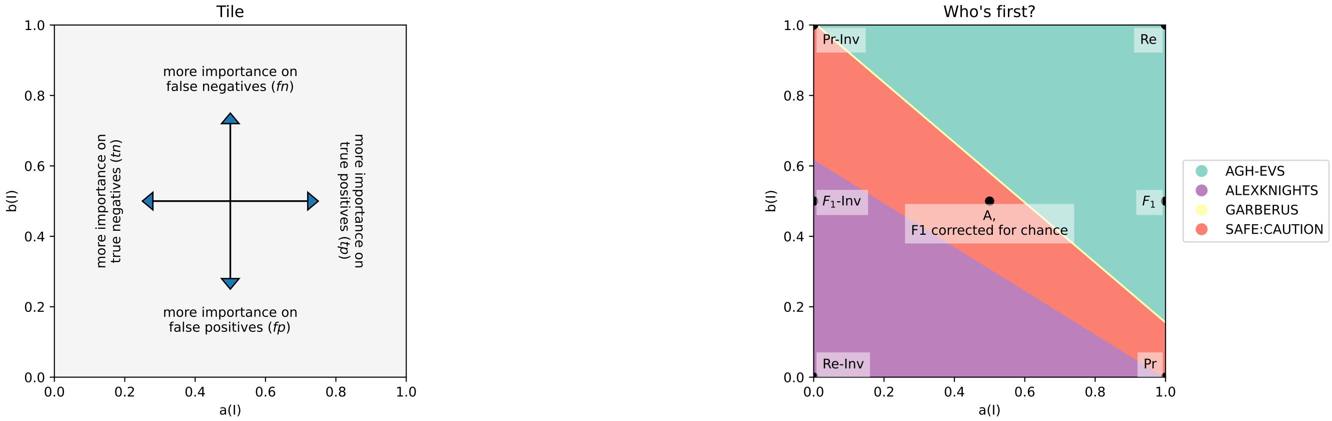


Figure 3. The Tile with annotations showing how to interpret it and the Tile with the best classifier.

It appears that all teams without exception have a confusion matrix very close to that of a no-skill classifier (flipping a coin, random guess): all balanced accuracy values are close to 0.5, as evident from the rising diagonal in ROC. For some participants, we can note a performance below no-skills. To isolate predictive merit from random chance, we applied the correction for chance [20]:

$$F_{1,\text{corr}} = \frac{F_1 - F_1 \circ \text{no-skill}}{1 - F_1 \circ \text{no-skill}}, \tag{5}$$

which rescales the metric so that zero represents the no-skill baseline ($F_1 \circ$ no-skill). Under this correction, SAFE:CAUTION achieves the best, albeit weak, result (0.12). As demonstrated in [20], it is known that the performance ordering induced by the chance-corrected versions of the F_1 is on the Tile. Moreover, as the classes are balanced, it is at the center of the Tile. Therefore, as the method proposed by SAFE:CAUTION has the highest value, it should appear on the Entity Tile[1]. This is indeed the case, as seen in Fig. 3.

Now, if we add the no-skill classifiers predicting always the same class to the ranking, it can be seen on Fig. 4 that SAFE:CAUTION is the only one to appear. All the other methods have a performance similar to random for the task defined by the contest on the test set. This result demonstrates that there is still room for improvement in possible new editions of the competition.

7. Conclusion

This paper introduced the first contest on illegal waste dumping detection, supported by the release of the MIVIA-IWDD dataset, namely the first publicly available video collection capturing both static and dynamic illegal disposal behaviors with precise temporal annotations. By addressing the long-standing lack of open benchmarks and by testing the methods on a challenging private test set, the competition establishes a solid foundation for developing and comparing video-based illegal waste dumping detection systems under realistic surveillance conditions.

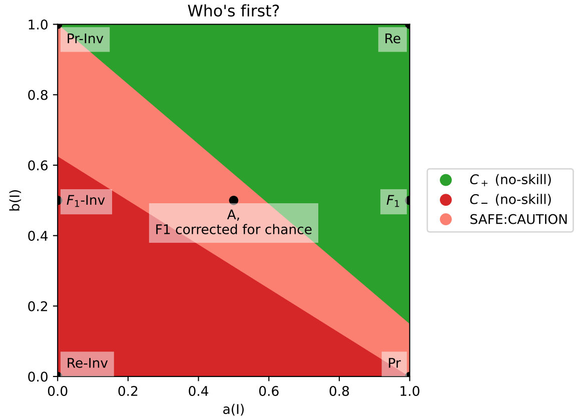


Figure 4. The Entity Tile with two additional no-skill classifiers.

The contest results reveal meaningful progress but also highlight the inherent difficulty of the task. While some methods demonstrated positive results and precise temporal localization, the overall performance is far from being acceptable for a real system, reflecting the challenges posed by subtle human actions, diverse environments, and the need for real-time operation. The heterogeneous strategies adopted by the ten participating team, ranging from spatio-temporal deep networks to multimodal and classical pipelines, underscore the multidimensional nature of the problem and the possibility to further improve the approaches with more robust temporal reasoning, motion modeling, and hard-negative discrimination.

[1] The Entity Tile, as proposed in [10], is a Tile on top of which we display the method for a given rank.

References

[1] Sajed Almorsy and Marwan Torki. X3d-based illegal waste dumping detection with temporal localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026. 4, 7

[2] Mudasir Bhat and Daw-tung Lin. Boundary-sensitive start-time estimation with onset-centric temporal detection for illegal waste dumping in surveillance video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026. 5, 7

[3] Aditi Bisht, Vishal Kamboj, Nitin Kamboj, Manisha Bharti, Kanchan Deoli Bahukahndi, and Himanshu Saini. Impact of solid waste dumping on soil quality and its potential risk on human health and environment. *Environmental Monitoring and Assessment*, 196(8):763, 2024. 1

[4] Rita Delussu and Lorenzo Putzu. A lightweight temporal detection framework for illegal waste dumping in real surveillance footage. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026. 5, 7

[5] Silvio Giancola, Anthony Cioppa, Bernard Ghanem, and Marc Van Droogenbroeck. *Deep Learning for Action Spotting in Association Football Videos*, page 427–459. WORLD SCIENTIFIC, 2025. 3

[6] Diego Gragnaniello, Antonio Greco, Carlo Sansone, and Bruno Vento. Onfire contest 2023: real-time fire detection on the edge. In *International Conference on Image Analysis and Processing*, pages 273–281. Springer, 2023. 1

[7] Diego Gragnaniello, Antonio Greco, Carlo Sansone, and Bruno Vento. Onfire 2023 contest: what did we learn about real time fire detection from cameras? *Journal of Ambient Intelligence and Humanized Computing*, 16(1), 2025.

[8] Antonio Greco and Bruno Vento. Par contest 2023: pedestrian attributes recognition with multi-task learning. In *International Conference on Computer Analysis of Images and Patterns*, pages 3–12. Springer, 2023.

[9] Antonio Greco, Alessia Saggese, Carlo Sansone, and Bruno Vento. An experimental evaluation of smart sensors for pedestrian attribute recognition using multi-task learning and vision language models. *Sensors*, 25(6):1736, 2025. 1

[10] Anaïs Halin, Sébastien Piérard, Anthony Cioppa, and Marc Van Droogenbroeck. A hitchhiker's guide to understanding performances of two-class classifiers, 2025. 8

[11] Nyayu Husni, Putri Sari, Ade Handayani, Tresna Dewi, Seyed Amin Hosseini Seno, Wahyu Caesarendra, Adam Glowacz, Krzysztof Oprzędkiewicz, and Maciej Sułowicz. Real-time littering activity monitoring based on image classification method. *Smart Cities*, 4:1496–1518, 2021. 2, 3

[12] Magzhan Kairanbay. Vision-language temporal analysis for illegal waste dumping detection in surveillance videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026. 4, 7

[13] SeongRok Kim, Donghun Ryu, Donghwan Hwang, Sungwoo Byeon, Minhyeok Lee, and Wonsuk Kim. K-fold ensemble of hiera and videomae for illegal waste dumping detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026. 5, 7

[14] Yeji Kim and Jeongho Cho. Aidm-strat: Augmented illegal dumping monitoring strategy through deep neural network-based spatial separation attention of garbage. *Sensors*, 22 (22), 2022. 2, 3

[15] Sriya Mahankali, Supreeth V Kabbin, Spoorti Nidagundi, and Ramamoorthy Srinath. Identification of illegal garbage dumping with video analytics. In *ICACCI*, 2018. 2, 3

[16] Sriram Reddy Mandhati, N. Lakmal Deshapriya, Chatura Lavanga Mendis, Kavinda Gunasekara, Frank Yrle, Angsana Chaksan, and Sujit Sanjeev. plitterstreet: Street level plastic litter detection and mapping, 2024. 2, 3

[17] Claudio Marrocco, Alessandro Bria, Francesco Tortorella, Sara Parrilli, Luca Cicala, Mariano Focareta, Giuseppe Meoli, and Mario Molinara. Illegal microdumps detection in multi-mission satellite images with deep neural network and transfer learning approach. *IEEE Access*, 12, 2024. 2, 3

[18] Alfredo Mazza, Prisco Piscitelli, Cosimo Neglia, Giulia Della Rosa, and Leopoldo Iannuzzi. Illegal dumping of toxic waste and its effect on human health in Campania, Italy. *International Journal of Environmental Research and Public Health*, 12(6):6818–6831, 2015. 1

[19] Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. Foundations of the theory of performance-based ranking. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14293–14302, 2024. 7

[20] Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. The tile: A 2d map of ranking scores for two-class classification, 2024. 7, 8

[21] Sébastien Piérard, Adrien Deliège, and Marc Van Droogenbroeck. What is the optimal ranking score between precision and recall? we can always find it and it is rarely F_1, 2025. 6

[22] Pedro F Proença and Pedro Simões. Taco: Trash annotations in context for litter detection, 2020. 2, 3

[23] Vincenzo Scarrica, Alessio Pierluigi Placitelli, and Antonino Staiano. Garberus: Three-headed video classifier to guard against illegal waste dumps. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026. 4, 7

[24] Ayesha Siddiqua, John N Hahladakis, and Wadha Ahmed KA Al-Attiya. An overview of the environmental pollution and health effects associated with waste landfilling and open dumping. *Environmental Science and Pollution Research*, 29(39):58514–58536, 2022. 1

[25] Rocio Nahime Torres and Piero Fraternali. Aerialwaste dataset for landfill discovery in aerial and satellite images. *Scientific Data*, 10(1):63, 2023. 2, 3

[26] Maria Triassi, Rossella Alfano, Maddalena Illario, Antonio Nardone, Oreste Caporale, and Paolo Montuori. Environmental pollution from illegal waste disposal and health effects: A review on the "triangle of death". *International Journal of Environmental research and public health*, 12(2): 1216–1236, 2015. 1

[27] Magdalena Daria Vaverková, Alžbeta Maxianová, Jan Winkler, Dana Adamcová, and Anna Podlasek. Environmental consequences and the role of illegal waste dumps and their impact on land degradation. *Land Use Policy*, 89:104234, 2019. 1

[28] Dasong Yu, Jungeun Yoon, and Youngjae Lee. Detection and register of illegal garbage dumping action using the consecutive processing and embedded-nas. In *2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2024. 2, 3

[29] Kimin Yun, Yongjin Kwon, Sungchan Oh, Jinyoung Moon, and Jongyoul Park. Vision-based garbage dumping action detection for real-world surveillance platform. *ETRI Journal*, 41(4):494–505, 2019. 2, 3

[30] Andrei-Daniel Zbuce, Lukas Laza, Andrei Ticarat, Elisa Moisi, and Daniela Popescu. Hybrid temporal-spatial novelty detection for illegal waste dumping in surveillance video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026. 5, 7