Journal of Scientific Reports 2026, 15(1), 103-127
ISSN: 2708-7085 (Online) & 3079-9317 (Print)
Journal: ijsab.com/jsr

Human Activity Classification for Electricity Infrastructure Protection in Rwanda

Ntambara Etienne [1*], Rukundo Simeon [2], Nkurunziza Egide [1], & Ingabire Daria [1]

[1] ICT Department, Rwanda Polytechnic Huye College, Huye, Rwanda.
[2] Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore.
* Corresponding author: Ntambara Etienne and ntambaraienne94@gmail.com

Abstract

The theft and vandalism of REG electricity infrastructure can disrupt essential services, raise maintenance costs, and pose safety hazards. However, no prior work has established a leakage-controlled, balanced image classification benchmark for activity cues relevant to this operational context. This study developed and tested an AI-based human activity recognition system as an early-warning tool to protect Rwanda Energy Group (REG) electricity infrastructure. We assembled two public Kaggle image datasets, Wall Climbing ALEE and Climbing V2, into six classes: climbing, cutting, normal, opening, suspicious, and vendors. The final dataset included 23,647 images: 19,554 for training, 1,814 for validation, and 2,279 for testing. Oversampling during training reduced the imbalance ratio from 13.751 to 1.000, while maintaining validation and test distributions. We fine-tuned a pretrained YOLO11m model with 224×224 pixel inputs, applying augmentation, AdamW optimization, dropout, cosine learning-rate scheduling, and early stopping. On the untouched test set, the model achieved 99.74% top-1 accuracy, 99.12% balanced accuracy, 99.40% macro F1-score, 0.9998 macro ROC-AUC, and 0.9971 macro average precision; it misclassified only six images out of 2,279. Suspicious activity detection remains a primary challenge, with 95.95% recall and 97.93% F1-score. GPU benchmarking on an NVIDIA Tesla T4 showed a mean latency of 6.64 ms and throughput of about 150.50 images/sec. These results support deploying the model as a high-performance visual perception layer, though operational use will require temporal modeling, sensor fusion, field validation, human oversight, and privacy safeguards.

Keywords: *Human activity recognition, CCTV, YOLO11, Electricity infrastructure security, Early warning, Rwanda Energy Group, REG.*

ARTICLE INFO
Research paper
Received: 20 July 2026
Accepted: 23 September 2026
Published: 25 September2026
DOI: 10.58970/JSR.1242

CITATION
Ntambara, E., Rukundo, S., Nkurunziza, E., & Ingabire, D. (2026). Human Activity Classification for Electricity Infrastructure Protection in Rwanda, *Journal of Scientific Reports*, 15(1), 103-127.

1. Introduction

Rwanda Energy Group (REG) operates electricity generation, transmission, and distribution assets, and protecting them is essential to reliable electricity service. Theft and vandalism of REG electricity infrastructure can interrupt supply, increase maintenance and replacement costs, reduce revenue, and create direct safety risks. Reported targets include electricity cables, pylons, transformers, meters, cross-arms, distribution equipment, and other network components. Recent Rwandan institutional reports demonstrate that this is not only a theoretical security problem but also an active challenge to REG's electricity infrastructure protection. Rwanda Energy Group (REG) has reported electricity losses associated with theft, including

approximately 6.5% commercial losses in one period and an estimated Frw 1.9 billion in electricity stolen by 28 customers identified during inspection operations (Rwanda Energy Group, 2020). REG also reported that theft and vandalism can cause outages, voltage drops, safety risks, and revenue losses. A separate 2019 report on REG cited an annual power loss of approximately Frw 19 billion. It stated that 44 MW, or about 19% of distributed power at the time, was lost through theft and technical losses (Buningwire, 2019). These figures are historical and are used here to show the scale of the problem rather than as current national estimates. In Rwanda, incidents of infrastructure vandalism have involved thieves, transporters, and scrap-metal buyers, with both electrical and communication assets being affected (Rwanda Energy Group, 2021; Rwanda National Police, 2020, 2021a). Arrests have been the result of community reports, which show that automated detection should complement rather than take the place of human observation (Rwanda National Police, 2021b). Recently, police operations succeeded in recovering meters, connectors, fuses, and cables, and recorded 829 cases relating to energy lines over a period of nine months (Rwanda National Police, 2023a). The existence of these patterns indicates that it is possible to identify physical interactions before or during the damage. Standard protection measures include the use of physical barriers, guards, patrols, lighting, alarms, community reporting, and closed-circuit television (CCTV). While CCTV offers continuous visual coverage together with recorded evidence, conventional monitoring places a heavy demand on human operators when it comes to interpretation. If a large number of cameras are operating at the same time, operators may fail to notice rare suspicious events, particularly when normal activity is the main feature of the video feed. Human activity recognition (HAR) provides a complementary method by automatically assigning observations to different activity categories. According to recent surveys, HAR has evolved from the use of handcrafted motion features to include convolutional, recurrent, graph-based, transformer-based, and multimodal deep-learning approaches (Dutta et al., 2025; Nguyen et al., 2023). Video-based HAR has also become more relevant to surveillance since actions usually have to be interpreted in light of the scene context and their temporal development (Surek et al., 2023). However, there is an important difference between image-level classification and event understanding. While a single image may offer useful evidence of what a person appears to be doing, it cannot reliably indicate whether there was intent or authorization, or whether theft has actually occurred. The recent literature on action recognition distinguishes among action classification, temporal localization, online action detection, and action anticipation (Heo et al., 2026). Similarly, research into surveillance anomaly detection emphasizes temporal information, since unusual events are typically defined by how a scene changes over time rather than by a single frame (Sultani et al., 2018; Ul Amin et al., 2024). The study now regards image classification as the first perceptual stage in a future system that involves CCTV, AI, sensors, and a human response. Its contribution consists of systematically curating two publicly available image datasets into a six-class taxonomy with an infrastructure focus, exercising explicit control over class imbalance and cross-split duplicates, fine-tuning the YOLO11m model for rapid classification, and carrying out a detailed assessment of both the predictive and operational metrics. The study is not intended to produce a universal surveillance classifier applicable to all public infrastructure; rather, it is aimed at developing a visually perceivable component that can be measured technically for use in a future system for protecting REG's electricity infrastructure. This clarification of the scope serves to determine the direction of the work. The experimental classifier is evaluated using an image benchmark that has been curated for research purposes. At the same time, REG provides the operational case study context, which aids in defining the problem, establishing the activity taxonomy, and setting the requirements for deployment. As a result, the study identifies three different levels of evidence: evidence from the curated images on the benchmark, evidence from reported incidents relating to REG, and deployment recommendations derived from the relationship between the two. It results in four important contributions. First, the study brings together two different public image

sources into a six-category activity classification system: climbing, cutting, normal, opening, suspicious, and vendors, based on visual clues relevant to protecting electricity infrastructure. The second contribution is that it introduces specific data-integrity procedures, for example, by detecting duplicates of exact content and by using oversampling only during training, in order to make sure that the performance on the test set is not affected by cross-split leakage or by artificially balanced datasets. The third is that it assesses YOLO11m using a range of metrics in addition to accuracy, such as class-balanced scores, ranking metrics, confidence behavior, and inference latency. The fourth is that it offers a deployment route suited to REG; this route combines frame-level classification with temporal evidence, physical sensors, asset data, and human verification. The study consequently asks the following research questions: RQ1: can a carefully selected six-class dataset serve as an appropriate visual activity taxonomy for early warning for REG electricity infrastructure? RQ2**:** Does simply balancing the classes during training adequately deal with severe imbalance while at the same time keeping the validation and test distributions unbiased? RQ3**:** Is it possible for YOLO11m to accurately tell the difference between climbing, cutting, normal, opening, suspicious, and vendor activities? RQ4**:** What is the model's performance for each class, especially for the suspicious class, which is operationally important? RQ5**:** Does the classifier offer adequate ranking quality and confidence behavior for use in an early-warning situation? RQ6**:** Is the trained model fast enough in terms of computing power to enable real-time or near-real-time CCTV analysis? The questions move from concerns about the validity of the data to those regarding predictive performance and computational feasibility; the deployment architecture is given later as a design implication for future field validation and not as an empirically answered research question. This research is organized as follows: Section 2 reviews the relevant literature and research gap; Section 3 presents the materials and methods; Section 4 reports the experimental results; Section 5 discusses the findings and their implications for REG electricity-infrastructure protection; and Section 6 concludes the study with its limitations and future research directions.

2. Literature Review

Human activity recognition

The identification of human actions represents a significant challenge in the fields of computer vision and machine learning, with the aim of detecting such actions from observational data. The more recent literature points to a fast transition from handcrafted features to sophisticated deep learning approaches, such as those based on spatial, temporal, transformer-based, and multimodal methods (Alomar et al., 2025; Dutta et al., 2025; Sedaghati et al., 2025). At present, HAR systems make use of a variety of data sources, such as RGB videos, depth sensors, skeletal models, wearable devices, information from Wi-Fi channels, radar, or combinations of these different modalities. In their review, Dutta et al. (2025) look at deep learning-based applications of HAR, the datasets used, the types of models employed, and the challenges that still remain concerning deployment and model generalization. Skeleton-based HAR offers a useful contrast to RGB classification. Nguyen et al. (2023) looked at techniques that model the spatial and temporal relationships between human joints, covering recurrent, convolutional, and graph-based architectures. Although these methods allow motion to be represented explicitly, they do so at the cost of requiring reliable pose or skeleton data. In this study, however, RGB images are used since the existing CCTV infrastructure naturally provides visual observations and since the first aim is to develop a lightweight perception layer rather than achieve full pose-based action understanding. Further studies in the field of video-based human activity recognition (HAR) show that having strong spatial representations by themselves is not enough for many actions. Surek et al. (2023) examined deep-learning methods for video HAR and demonstrated that architectures which combine spatial representation with temporal information remain important. This point is directly relevant to infrastructure protection since opening a cabinet,

climbing a barrier, cutting a component, and staying within a restricted area are all processes, not merely static states. Surveys carried out in recent years have extended the taxonomy of action recognition from classification to include temporal action localization, spatio-temporal localization, temporal segmentation, online action detection, and action anticipation (Heo et al., 2026). The current study deliberately focuses on image-level classification only. By adopting this more limited scope, the study is able to set up a high-speed baseline which can afterward be combined with tracking, temporal aggregation, or action-recognition transformers.

Deep learning for surveillance and anomaly detection

Smart surveillance systems are now employing deep learning in order to decrease the amount of video that has to be examined by people themselves. Sultani et al. (2018) showed how weakly supervised anomaly detection can be achieved in long, untrimmed surveillance videos, highlighting the advantages of learning from both normal and abnormal events at the video level. More recent studies have made use of 3D convolutions and LSTM modules in order to capture both short- and long-range spatiotemporal information and to cut down the number of false alarms in surveillance anomaly detection (Ul Amin et al., 2024). Research shows that surveillance systems should not base their operational decisions regarding the progression of an event solely on a single frame. Investigations into anomaly detection highlight the significance of scene context, temporal cues, and the fact that labeled abnormal events are limited (Duja et al., 2024; Wastupranata et al., 2024). For instance, a person close to a secure site might be a worker, a vendor, a pedestrian, or an intruder; the same appearance can have different meanings depending on authorization, location, how long they stay, and their interaction with the infrastructure. Hence, a frame classifier works most effectively as a source of evidence rather than as a system that makes independent decisions. Recent research into lightweight HAR also highlights the conflict between recognition performance and computational cost. AlMuhaideb et al. (2024) looked at methods aimed at maintaining a useful level of recognition capability while at the same time reducing computational demands, thus underlining the need for model design that takes deployment into account. In this study, therefore, we specifically assess inference latency rather than giving only accuracy figures.

Class imbalance and data augmentation

Security datasets usually include a great deal of normal activity and only a small number of events that are relevant to security. When a classifier is trained directly with such an imbalance, the overall accuracy may mask a weak recall for the minority class. Khan et al. (2024) looked at and assessed various methods for dealing with class imbalance and found that resampling and augmentation can provide effective and computationally efficient solutions as alternatives to more complicated synthetic-generation methods. It has also been suggested that class-balanced learning should be used as an alternative to frequency-driven empirical risk minimization (Cui et al., 2019). Oversampling is only one of the methods available for dealing with imbalances; techniques such as class-weighted cross-entropy impose a heavier penalty on errors relating to the rare classes, while class-balanced loss sets weights according to the effective sample size. Focal loss decreases the impact of easy examples so that the model can concentrate on more difficult cases (Cui et al., 2019; Lin et al., 2017). These approaches avoid the need for direct data duplication but do introduce weighting or focusing parameters, which do not necessarily lead to an increase in appearance diversity. In this study, we employed transparent oversampling that is used only during training together with online augmentation, thus retaining the standard classification objective and making class exposure explicit. We acknowledge that a direct comparison with focal and class-balanced losses will be left for future work and do not regard it as an advantage claimed here. Data augmentation increases the variety of the training examples and is able to reduce overfitting in cases where the amount of available data is small. Zeng et al.

(2024) looked at modern image-augmentation techniques and examined their contribution to improved generalization. Other comprehensive reviews also identify geometric, photometric, mixing, and synthetic augmentation approaches (Kumar et al., 2024; Mumuni et al., 2024). In the present study, we carry out augmentation during the training process, although we limit oversampling to the training subset. This point is important since balancing the validation and test data would alter the empirical distribution that is used for the final evaluation. Transfer learning provides yet another means of reducing the requirement for large amounts of task-specific data. Networks that have already been pretrained pick up general low- and mid-level features that can be adjusted to new classification tasks. The model currently in use begins with a pre-trained YOLO11m classification checkpoint and is then fine-tuned on six classes that have been carefully chosen. This approach is in line with wider research showing that pretrained deep representations can be effectively adapted to downstream visual tasks (He et al., 2016).

Edge computing and multimodal sensing

Edge computing can be of benefit to CCTV analytics since video streams are both bandwidth-intensive and security-sensitive. Wazwaz et al. (2024) looked at the distribution of HAR across smart devices, IoT edges, and cloud computing, highlighting the advantages regarding latency and communication that come with local or distributed processing. When it comes to protecting infrastructure, an edge model is able to process the relevant frames close to the camera and send only the events or reduced metadata to the control center. Multimodal sensing provides information that cameras alone cannot. Varga (2025) showed how decision fusion can be achieved using Wi-Fi channel-state information, emphasizing the advantages of combining several sensing methods. In the context of infrastructure security, visual data can be supplemented by fence-vibration sensors, magnetic door contacts, electrical measurements, acoustic sensors, PIR detectors, geofencing, and details regarding access control. This kind of multimodal strategy is particularly useful for identifying suspicious activities since alerts supported by multiple kinds of physical evidence are more reliable than those based solely on visual inspection. Moreover, research into wireless sensor networks and deep learning shows that distributed sensing systems are important (El-Adawi et al., 2024). Models like I3D and SlowFast demonstrate the importance of temporal information for understanding actions (Carreira & Zisserman, 2017; Feichtenhofer et al., 2019), and non-local and transformer-based models offer ways of capturing long-range dependencies (Liu et al., 2022; Wang et al., 2018). This is why the study has decided to combine training-only balancing, image augmentation, fast inference, and a future sensor-fusion layer. In all these areas, a clear design principle can be identified: the fundamental security unit is not merely a classification label but a sequence of evidence that is assessed in the context in which it is found. RGB HAR provides the visual aspect; techniques for dealing with class imbalance assist in handling the rarity of observations that are relevant to security; edge computing lowers the cost and delay associated with continuous video processing; and multimodal sensing supplies data that images on their own cannot capture reliably. Rather than asserting that one neural classifier is capable of tackling the whole security problem, this study combines these various elements in a step-by-step approach. By doing so, the research brings together a number of well-established methodological features that have been adapted for a particular REG operational scenario. The reason that this six-class classification includes both potentially harmful and legitimate activities is made clear here. It would be unrealistic and result in a large number of false alarms if a security system treated anyone who was near an electrical asset as suspicious. Conversely, if the taxonomy had only normal and theft categories, it would fail to cover intermediate cases such as climbing or opening. Vendors are also included for a specific reason: legitimate commercial activities can appear similar to suspicious behaviors, and there are reported instances of infrastructure cases in which resale channels have become relevant during investigations. This taxonomy is designed to reflect operationally

significant ambiguity and does not assume that simply being present visually means there is criminal intent. The literature suggests that evaluation should account for deployment risk. Although overall accuracy is useful, it can hide problems that occur in the minority classes. ROC-AUC and average precision measure ranking performance but do not verify temporal event detection. Even if the classification accuracy is high, the confidence scores may still be overconfident. Latency figures show that the system is computationally feasible but fail to address the whole CCTV-to-alert process. The fact that there are these differences means that the multi-dimensional evaluation employed here is justified and provides a careful basis for interpreting the results. This technical evidence is also part of a wider socio-technical debate. It is not possible to assess surveillance performance merely on the basis of model accuracy, since the way in which alerts are interpreted and dealt with is influenced by the operators, the institutional procedures, the authorization data, and the communities concerned (Baxter & Sommerville, 2011). Moreover, the theory of contextual integrity stresses that privacy hinges on whether the flows of data are consistent with the situation, the people involved, the purposes, and the rules governing transmission, not just on the collection of data (Nissenbaum, 2004). With regard to the proposed REG use case, these perspectives justify measures such as purpose limitation, human oversight, audit logs, role-based access, data retention limits, and procedures for appeal or correction. They also explain why the classifier is regarded as a tool for decision support rather than as an autonomous accusation tool.

Design implications for electricity-infrastructure protection

The literature as a whole indicates that an infrastructure-security HAR system has to satisfy four important criteria. It should be able to visually differentiate between various activity states with a sufficient degree of accuracy. It also needs to be computationally efficient so as to be able to deal with continuous CCTV feeds, these feeds possibly being processed at a number of different locations. The system must furthermore specifically take account of the problem of class imbalance and the rarity of real harmful events. Finally, it should clearly separate the perception phase from the decision-making phase. These criteria lend support to a layered architecture in which a fast visual classifier produces possible evidence and subsequent modules—making use of temporal and multimodal data—decide whether or not an alert is required.

Research gap

Even though there have been considerable developments in HAR, surveillance anomaly detection, class imbalance management, and edge AI, few studies have clearly connected a curated taxonomy of electricity-security activities with the operational early-warning requirements for REG. Standard benchmark datasets generally do not reflect the actual conditions found at REG cameras, such as different viewpoints, lighting conditions, weather, infrastructure designs, maintenance procedures, authorized technicians, and the usual activities surrounding REG sites. Furthermore, a high level of benchmark accuracy does not necessarily indicate that the system is ready to be put into use. The research gap in this area goes beyond classification accuracy to include the need for a relevant activity taxonomy, leakage control, imbalance handling, predictive assessment, latency evaluation, and the various integration pathways involving CCTV, sensors, and human supervision.

Hypotheses Development

Prior studies have shown that pretrained deep representations can be effectively adapted to downstream visual tasks through transfer learning (He et al., 2016), that resampling and augmentation provide effective solutions to class imbalance (Khan et al., 2024), that lightweight HAR methods must balance recognition performance against computational cost (AlMuhaideb et al., 2024), and that confidence behavior and ranking quality are essential for early-warning

systems (Sultani et al., 2018; Ul Amin et al., 2024). Therefore, the following hypotheses have been developed:

Overall Classification Performance

Prior studies have shown that pretrained deep representations can be effectively adapted to downstream visual tasks through transfer learning (He et al., 2016), and that careful data curation with training-only oversampling preserves the integrity of the test set (Khan et al., 2024). Therefore, the following hypothesis has been developed:

H1: *The six-class YOLO11m classifier, having been fine-tuned, will attain a macro F1-score of 0.90 or more on the test set, which has not been touched.*

Minority Class Recall

Prior studies have shown that security datasets typically contain abundant normal activity and rare security-relevant events, and that resampling and augmentation can effectively address class imbalance without altering the evaluation distribution (Khan et al., 2024). Therefore, the following hypothesis has been developed:

H2: *Following balancing based only on training data, the minority suspicious class will have a test recall of at least 0.90 without having the validation or test sets balanced.*

Inference Latency

Prior studies have shown that lightweight HAR methods must balance recognition performance against computational cost, and that edge computing reduces latency for continuous video processing (AlMuhaideb et al., 2024; Wazwaz et al., 2024). Therefore, the following hypothesis has been developed:

H3: *The average model inference latency on the specified GPU benchmark will stay under 40 ms per image, which corresponds to the per-frame budget of 25 frames per second.*

Selective Acceptance and Confidence Behavior

Prior studies have shown that confidence behavior and ranking quality are essential for early-warning systems, and that a frame classifier works most effectively as a source of evidence rather than as an independent decision-maker (Sultani et al., 2018; Ul Amin et al., 2024). Therefore, the following hypothesis has been developed:

H4: *If a selective acceptance is applied at a confidence threshold of 0.95, the accuracy of the accepted sample will be greater than that obtained by accepting all the predictions, even though it is not anticipated to eliminate every false negative that has a security relevance.*

Rwandan operational evidence and system requirements

The REG and RNP reports highlight three key design requirements. Since incidents may involve direct actors, transporters, and resale intermediaries, the system must distinguish between potentially harmful actions and legitimate activities involving presence and vendors (Rwanda Energy Group, 2021; Rwanda National Police, 2020). Because assets such as cables, meters, transformers, and distribution equipment are at risk, visual evidence should be complemented by physical sensors suitable for each asset type (Rwanda National Police, 2023a). Finally, the involvement of community reporting and inter-agency coordination underscores the need for a well-defined human response pathway for alerts (Rwanda National Police, 2021b, 2024).

3. Materials and Methods

Research design and workflow

The study employed an experimental machine-learning approach that included acquiring a public dataset, curating the classes, verifying image integrity, detecting exact-content duplicates,

removing cross-split leakage, and performing various training-related procedures such as balancing, augmentation, transfer learning, selecting the model based on validation results, and independent testing along with evaluation confidence analysis and inference latency measurement. The system was implemented in Python using PyTorch, Torchvision, Ultralytics, NumPy, Pandas, scikit-learn, PIL, and Matplotlib within a Kaggle GPU environment.

Dataset sources and curation

Our study used two public datasets from Kaggle. The first was the Wall Climbing ALEE / Wall Security Dataset (Qureshi, 2026). The second was Climbing V2 (ngcduy09on, 2024). The research did not treat the source datasets as representing an already validated six-class Rwandan CCTV classification system. Rather, the researchers organized the images into six operational categories relevant to infrastructure protection: climbing, cutting, normal, opening, suspicious, and vendors. Climbing means people scaling walls, fences, or other barriers; cutting refers to activities involving tampering; normal describes ordinary activity that has no connection with security threats; opening involves interaction with doors, gates, cabinets, or access points; suspicious denotes behavior that is considered operationally suspicious even though it has not been clearly assigned to any particular action; and vendors include legitimate commercial or roadside activities. It is important to include both vendors and normal activity since an effective early-warning system must be able to tell the difference between legitimate presence and potentially dangerous behavior. The researchers also reduced the image size to match the standard model input of 224 × 224 pixels and employed training augmentation in order to enhance robustness. Since the original datasets are public and diverse, the resulting six-class dataset is clearly regarded as a research-curated benchmark and not as a nationally representative CCTV dataset.

Dataset integrity, leakage control, and input preparation

Prior to the training process, we examined the clarity of the images. Each image that was clear was given an MD5 hash so that exact duplicates could be identified. There were seven different hashes among the various splits, which showed that ten images had been removed from the lower-priority splits. Crucially, none of the duplicate hashes had conflicting class labels. We kept the original order of the images—that is, the order for training, validation, and test—so that no identical images would end up in both the training and the evaluation sets. This careful approach helped to avoid artificially inflated performance metrics. The final model input resolution was set at 224 by 224 pixels. When training was carried out, horizontal flipping was applied with a probability of 0.50, vertical flipping was not used, and random erasing was performed with a probability of 0.20. The pipeline also incorporated RandAugment and had a dropout rate of 0.25. The validation and test images were not subjected to oversampling. This method clearly distinguishes between three major interventions: curation, which establishes the target taxonomy; oversampling, which modifies the training frequencies; and augmentation, which increases the diversity of the training data.

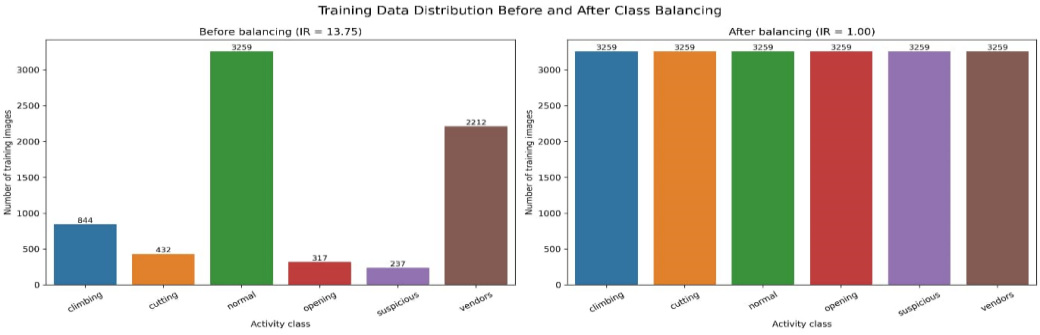

Figure 1: Training-set class distribution before and after balancing.

Figure 1, the training subset was oversampled, and as a result, the original imbalance ratio of 13.751 to 1.000 was reduced. Initially, the normal activity had 3,259 training images while suspicious activity had only 237. After the oversampling, which was carried out on the training data alone, each of the six classes had 3,259 images.

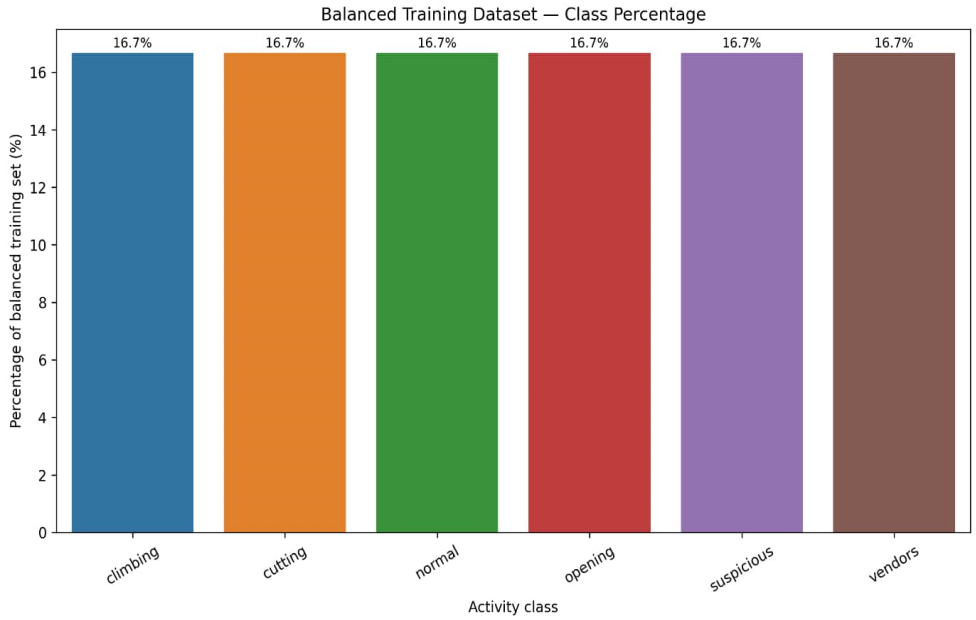

Figure 2: Percentage composition of the balanced training dataset.

Each activity class contributes approximately 16.7% of the 19,554-image training set, as illustrated in Figure 2. From Table 1 only the training partition was balanced.

Table 1: Final curated dataset after leakage cleaning.

Class	Train	Validation	Test	Total
Climbing	3259	210	264	3733
Cutting	3259	107	135	3501
Normal	3259	807	1016	5082
Opening	3259	79	99	3437
Suspicious	3259	59	74	3392
Vendors	3259	552	691	4502
Total	19554	1814	2279	23647

Model and training configuration

We selected YOLO11m-cls, a pretrained YOLO11m image-classification checkpoint (see table 2). We fine-tuned the model for six classes using 224 × 224-pixel inputs. The training configuration used an initial learning rate of 0.001, AdamW optimization, weight decay of 0.0005, batch size 32, a maximum of 80 epochs, early-stopping patience of 15 epochs, dropout 0.25, horizontal flip 0.50, random erasing 0.20, RandAugment, cosine learning-rate scheduling, random seed 42, and an NVIDIA Tesla T4 GPU. Training terminated after 66 recorded epochs under the early-stopping policy (Ultralytics, 2025).

Table 2: Principal model-training configuration.

Setting	Value
Model	YOLO11m-cls
Input size	224 224
Optimizer	AdamW
Initial learning rate	0.001
Weight decay	0.0005
Batch size	32
Maximum epochs	80
Early stopping patience	15 epochs
Dropout	0.25
Horizontal flip	0.50
Random erasing	0.20
Augmentation	RandAugment
Learning-rate schedule	Cosine
Random seed	42
GPU	NVIDIA Tesla T4

Operational interpretation of the six classes

The six categories were established as types of visual activity rather than as legal classifications. The category 'Climbing' includes cases of people climbing walls, fences, poles, or barriers; 'Cutting' includes any visible use of cutting actions or tools which might indicate tampering; 'Opening' refers to interactions with doors, gates, cabinets, or other kinds of access points; and 'Normal' includes everyday activities that do not correspond to the specific behaviors in question. 'Vendors' stands for legitimate commercial or roadside activity which can take place near infrastructure, while 'Suspicious' denotes behavior which is visually concerning and cannot be reduced to a single, mechanically defined action. This final category is inherently more dependent on context than the others and is therefore regarded as a high-priority but non-conclusive indicator. In the REG case study, these labels are meant to deal with a specific question regarding perception: what visual state is there in a frame? They do not address whether a person is authorized, whether a component belongs to REG, whether material has been removed, or whether a crime has taken place. To answer those questions, it is necessary to have temporal information, asset metadata, access-control records, or human investigation. By keeping this distinction, it is avoided that the classifier appears to be an independent enforcement mechanism. Following curation, the data was kept in separate training, validation, and test sets. To detect exact-content duplicates between the different partitions, we used MD5 hashes. In the case of finding duplicate files across the splits, we kept the first instance according to the following priority: training, then validation, then test. The aim of this procedure was to avoid a rather simple problem in which the model could come across the same image many times during both the optimization and the evaluation phases. It is important to note that we did not rebalance the final distributions of the validation and test sets; instead, we kept their natural class frequencies so that we could assess performance under the actual distribution resulting from the curation process. The original training data were substantially imbalanced, with an imbalance ratio of

13.751. We therefore applied oversampling only to the training partition until each class had 3,259 training images, reducing the training imbalance ratio to 1.000. We did not oversample validation or test images. This design avoids changing the evaluation target while allowing the optimizer to encounter minority-class examples more frequently. Augmentation was also confined to the training stage. In order to enhance the appearance diversity, we used horizontal flipping (with a probability of 0.50), random erasing (with a probability of 0.20), and RandAugment. A dropout rate of 0.25 was employed during training. These operations were selected so as to improve the model's robustness against variation in pose, framing, clothing, background, and partial occlusion without causing a synthetic evaluation distribution. We selected YOLO11m-cls as our primary model since the study required a fast image-level classifier that could then act as an edge perception layer. The model was initialized with pre-trained weights and subsequently fine-tuned on the six specified classes. Training was performed on inputs of size 224 × 224 using AdamW, with an initial learning rate of 0.001, a weight decay of 0.0005, a batch size of 32, cosine learning-rate scheduling, a dropout rate of 0.25, and for up to 80 epochs. Early stopping was applied with a patience value of 15 epochs so that training would halt if the validation performance stopped improving. A fixed random seed of 42 was used to improve reproducibility, and the experiments were carried out on an NVIDIA Tesla T4 GPU. The evaluation divides recognition quality from deployment feasibility; although accuracy reflects the overall success, balanced accuracy and macro F1 are employed in order to reduce the dominance of the majority class, ROC-AUC and average precision are used to assess ranking, confidence analysis is used to evaluate individual reviews, and latency is used to assess the model's computational performance. Nevertheless, no single one of these metrics can on its own establish end-to-end event detection, field generalization, or operational safety. Performance was assessed using top-1 accuracy, balanced accuracy, macro precision, macro recall, macro F1-score, weighted F1-score, the Matthews correlation coefficient (MCC), Cohen's kappa, one-vs-rest ROC-AUC, average precision, confusion matrices, a confidence analysis, a confidence-threshold analysis, and inference latency. The choice of balanced accuracy and the macro metrics was based on the fact that the test set remained imbalanced. When measuring inference latency, 100 test images were used following three warm-up predictions, and CUDA synchronization was performed both before and after the GPU inference in order to avoid the asynchronous execution time being underestimated.

4. Results

Training behaviour

The training process involved 66 epochs, with the final training loss amounting to approximately 0.00249 and the validation loss to about 0.00507, with the validation top-1 accuracy reaching 0.9989 and the top-5 accuracy coming to 1.0000. In carrying out the final evaluation, we chose to use the best checkpoint instead of assuming that the last one was the optimal one. Both losses decrease rapidly and then stabilize at low values.

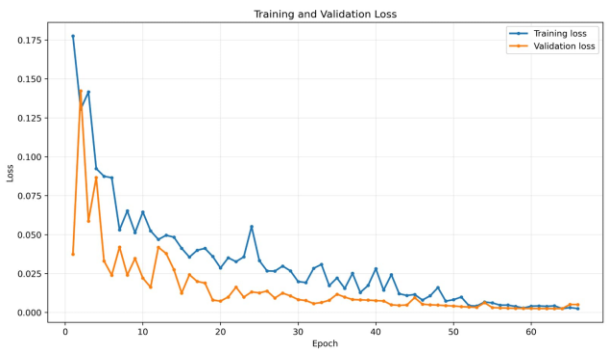

Figure 3: Training and validation loss across the recorded epochs.

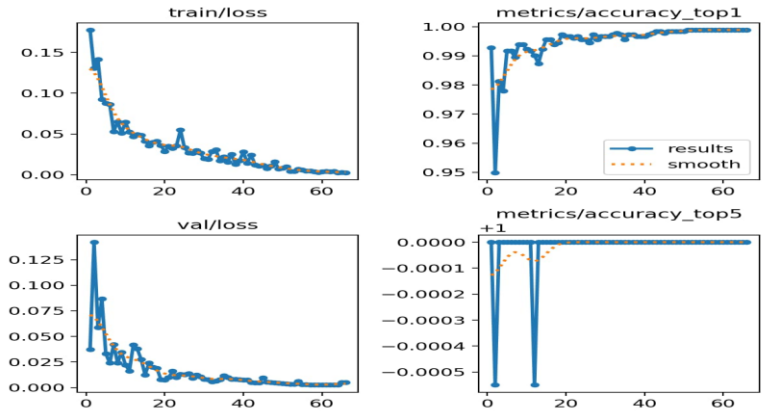

Figure 4: Ultralytics training log: loss and validation/top-1 performance across epochs. Validation top-1 accuracy approaches 1.00 while loss decreases toward zero.

Figure 3, presents the improvements appear at first, then level off gradually. Towards the end, the training loss stays below the validation loss as one would expect, with the validation curve remaining low. The small gap in the later stages shows that you need to choose the best validation checkpoint rather than the final epoch as the optimal model. model. Figure 4, the training-log view shows that validation top-1 accuracy tends toward 1.00 as the loss decreases. Although the top-5 curve is not useful for this six-class case when it reaches 1.00, it does show that the model consistently kept the correct class among its top-ranked predictions.

Dataset integrity and class-balance results

Table 3: Training-set balancing and oversampling factors.

Class	Original	Balanced	Factor
Climbing	844	3,259	3.86
Cutting	432	3,259	7.54
Normal	3,259	3,259	1.00
Opening	317	3,259	10.28
Suspicious	237	3,259	13.75
Vendors	2,212	3,259	1.47
Total	7,301	19,554	–

The audit of the dataset revealed two key points concerning the final metrics. At first, there were 11,404 readable image files in the source partitions before the cross-split duplicates had been removed. By using exact MD5 matches, 57 duplicate records were identified, all of which had 16 different hashes and were all from the normal class. Seven of the hashes were found in more than one partition, resulting in 10 validation and test files that contained images also present in the training set. We took these 10 files out of the subsequent partitions, which left 11,394 images before oversampling was carried out on the training data. The other duplicates were all within the training set and therefore did not have a direct effect on the validation or testing, but they did slightly decrease the diversity of the training data. The original training set was highly imbalanced, with 3,259 normal images (44.64%) and only 237 suspicious ones (3.25%). To balance the dataset, we oversampled each minority class to match the 3,259 normal images,

resulting in a total of 19,554 images (see table 3). The oversampling factors were 3.86 for climbing, 7.54 for cutting, 10.28 for opening, 13.75 for suspicious, and 1.47 for vendors. We applied the highest factor to the critical suspicious class. Since oversampling duplicates existing examples rather than introducing new, independent events, this approach enhances class exposure during training but does not increase the diversity of real-world situations.

Overall test performance

On the 2,279-image untouched test set, top-1 accuracy was 99.7367% and top-5 accuracy reached a perfect 100.00%. The independent measures showed a balanced accuracy of 99.1155%, a macro precision of 99.6952%, a macro recall of 99.1155%, a macro F1 score of 99.3975%, a weighted F1 score of 99.7355%, a MCC of 99.6182%, and a Cohen's kappa of 99.6180%. The test set was deliberately made imbalanced, containing 1,016 normal images and just 74 suspicious ones. The fact that the results were consistent across the accuracy, weighted F1, and class-balanced metrics shows that the high performance is not entirely the result of a bias towards the majority class. Nevertheless, since the number of suspicious images is small, this class is still statistically underrepresented and has to be interpreted on its own.

Per-class performance

Table 4: Per-class test performance.

Class	Precision	Recall	F1	Support
Climbing	99.25%	100.00%	99.62%	264
Cutting	99.26%	100.00%	99.63%	135
Normal	99.80%	99.90%	99.85%	1016
Opening	100.00%	98.99%	99.49%	99
Suspicious	100.00%	95.95%	97.93%	74
Vendors	99.86%	99.86%	99.86%	691

The model showed strong performance in all six categories, with climbing and cutting reaching 100% recall, normal achieving 99.90% recall, opening attaining 98.99%, and vendors 99.86% (see table 4). The main area of residual weakness was suspicious activity: although the precision was 100.00%, the recall was only 95.95%, and the F1 score was 97.93%. In practical terms, this indicates that the predictions classified as suspicious were very reliable on this test set, but some images that were actually suspicious were not detected.

Confusion matrix analysis

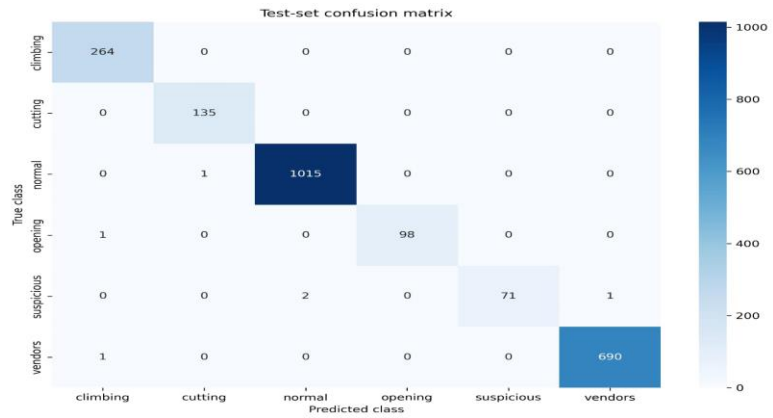

Figure 5: Test-set confusion matrix. Only six errors occurred among 2,279 test images.

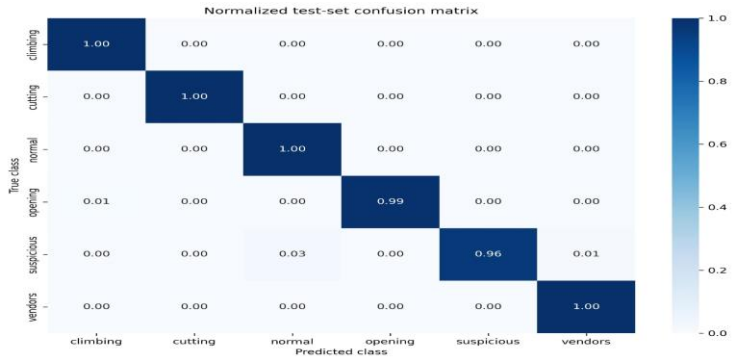

Figure 6: The confusion matrix for the test set, with the rows normalized. The recall for the suspicious class is about 0.96, while for the other classes it is at or near 1.00.

The confusion matrix in Figure 5, there is a strong concentration of diagonal errors. The six errors included two that were suspicious and normal, one that was suspicious and directed towards vendors, one that was normal and directed towards cutting, one that went from vendors to climbing, and one that went from opening to climbing. Since three of the six errors (50 percent) involved suspicious activity, even though suspicious activity made up only 3.25 percent of the test set, the remaining errors are therefore concentrated around the most operationally important class rather than being spread out over the entire taxonomy. Specifically, the errors going from suspicious to normal and from suspicious to vendors show that it is visually difficult to tell legitimate activity apart from security-related behavior in a single frame. Figure 6, by normalizing each row, the class-specific pattern is made clearer. The suspicious class still has about 96% correct recognition, with the rest split between normal and vendors. Meanwhile, the opening-to-climbing error shows that physical interaction with barriers can be visually ambiguous within a single frame.

ROC-AUC and precision–recall performance

Figure 7: Errors on the test set with high confidence.

The examples illustrate that visually ambiguous or context-dependent activities can have a high level of model confidence even when the label is wrong. The ROC-AUC values for the one-vs-rest approach varied from 0.9990 in the case of the suspicious class to 1.0000 for the four classes, giving a macro ROC-AUC of 0.9998. Average precision ranged from 0.9830 for the suspicious class to 1.0000 for climbing and vendors, resulting in a macro average precision of 0.9971. The lower average precision values for the suspicious class are in agreement with the confusion matrix and are particularly informative when considering the imbalanced nature of the test distribution:

accuracy alone fails to reflect the remaining difficulty involved in ranking the rare security-relevant class.

Confidence and threshold behavior

The mean confidence level for correct predictions was about 0.99915, and the median was 1.00000; for incorrect predictions, the mean confidence was about 0.90108, and the median was about 0.97990, the highest confidence level for an incorrect prediction being 0.99997. More significantly, all three of the missed suspicious cases were given extremely high confidence in their incorrect predictions: 0.99997, 0.99992, and 0.99985. Hence, the main security-related false negatives were not low-confidence borderline cases. A confidence value from a softmax-like classifier should therefore not be taken as a guaranteed probability that a prediction is correct. This is in line with the wider body of calibration literature, which demonstrates that even high-performing modern neural classifiers can still generate poorly calibrated confidence estimates (Guo et al., 2017). Figure 7, this is particularly important when it comes to deployment interpretation. Although the model may be very confident when it is in fact incorrect, a security policy should not cause an irreversible response based on a single confidence score; instead, it is necessary to combine the confidence level with temporal persistence, location, authorization status, and evidence from independent sensors. A further analysis based on selective prediction shows that thresholding can improve the accuracy of the predictions that are automatically accepted, although only slightly and with a significant security limitation. When a confidence threshold of 0.95 was used, 99.649% of the test samples were still accepted, and the accuracy for the accepted samples rose to 99.824%. Yet thresholding rejected only 0.351% of the samples and did not get rid of the high-confidence suspicious false negatives mentioned above (see table 5). Accordingly, the results indicate that confidence thresholding should be regarded as a mechanism for prioritizing reviews, not as a replacement for temporal or sensor-based verification.

Table 5: High-confidence test-set errors. The examples show that visually ambiguous or context-dependent activities can receive high model confidence despite an incorrect label.

Threshold	Coverage	Accepted accuracy	Rejected
0.50	100.000%	99.737%	0.000%
0.70	99.781%	99.780%	0.219%
0.90	99.737%	99.780%	0.263%
0.95	99.649%	99.824%	0.351%

Qualitative prediction examples

Figure 8: A high-confidence suspicious prediction is given as an example.

The visual evidence shows the need for context and temporal persistence before treating a prediction as an operational event. A correctly classified suspicious frame is shown in Figure 8. Nevertheless, the visual clue is contextual, not definite; it is the posture and proximity that indicate attention to the surrounding structure, whereas the image by itself does not show authorization, intent, or a harmful act that has already been carried out. The example thus supports the use of the suspicious label as a trigger for review, not as evidence of theft or vandalism.

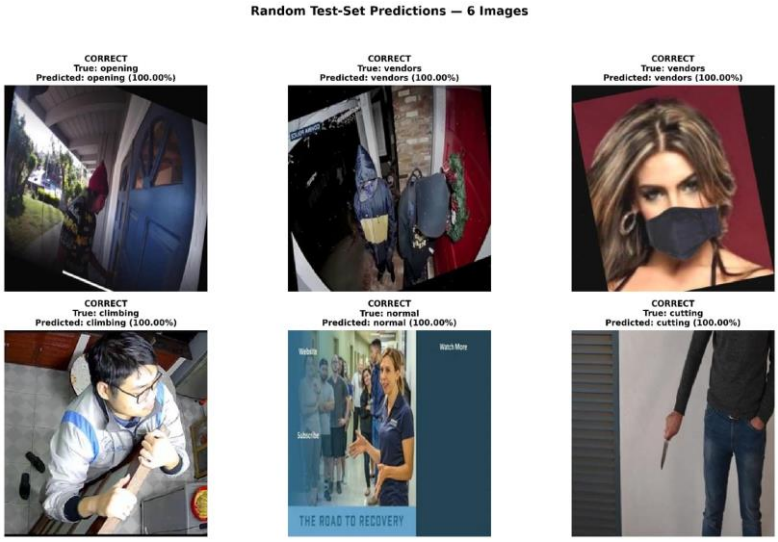

Figure 9: Random test-set predictions across the six activity classes.

The displayed examples were correctly classified with high confidence. Figure 9, shows correctly classified examples spanning legitimate and potentially security-relevant activity, supporting the multi-class rather than binary taxonomy. The montage is only a qualitative check, not independent evidence of performance. Its near-perfect confidence must also be read alongside Figure 7, where incorrect predictions received similarly high confidence; visual plausibility and confidence therefore cannot replace event-level verification.

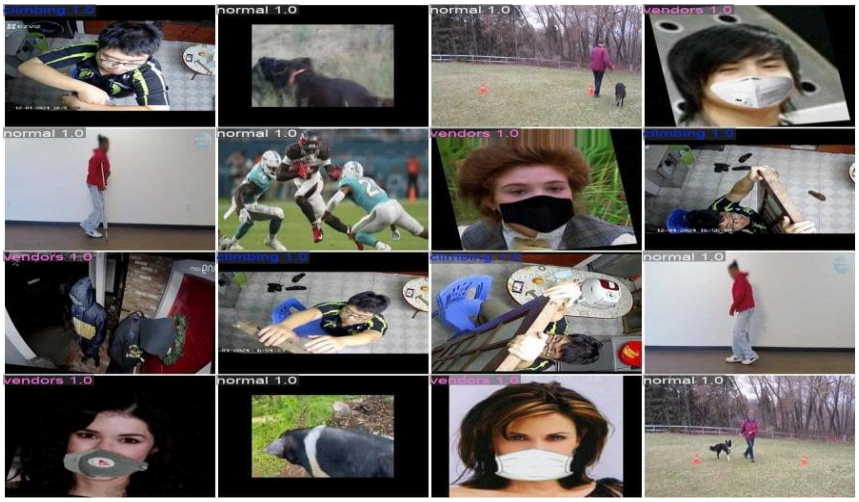

Figure 10: Additional qualitative test-set prediction montage.

The examples demonstrate variation in viewpoint, background, clothing, lighting, and activity appearance. The broader montage in Figure 10 shows that the curated dataset includes a great deal of visual variation and also explains why domain shift is still a major issue, since some of the

examples are CCTV-like while others differ greatly from the camera geometry and the environmental conditions to be expected in a Rwanda deployment.

Alignment with observed infrastructure incidents

The six-class classification agrees with the activity that has been reported regarding Rwandan infrastructure while still retaining important uncertainties. The actions of climbing, cutting, and opening mirror the visible interactions mentioned in the incident reports, and the normal and vendor categories stand for legitimate activity that is associated with commercial dealings and should not be automatically regarded as criminal (Rwanda Energy Group, 2021; Rwanda National Police, 2020, 2022, 2023a, 2024). Although this correlation enhances operational relevance, it does not mean that the public-image benchmark reflects the actual conditions in the field of the Rwanda Energy Group. The effects reported are outages, repair expenses, business disruption, and safety hazards; for this reason, the operational aim is the establishment of a procedure for dealing with an event rather than assigning responsibility to an individual (Rwanda Energy Group, 2020, 2021, 2025). Visual interactions that are repeated and supported by vibration, contact, electrical, or access-control signals should be given priority over a single suspicious frame. Historical loss figures should not be mixed up with the model's predictive metrics. The figure of Frw 19 billion per year for power loss and the amount of Frw 1.9 billion linked to the 28 cases of theft identified are institutional and media estimates based on particular time periods and inspection campaigns (Buningwire, 2019; Rwanda Energy Group, 2020). These figures show the practical significance of the problem; yet they do not give the true prevalence rate for the six activity categories, nor do they offer a direct estimate of the economic benefit of using this classifier.

Inference speed

For the NVIDIA Tesla T4, 100 test images were benchmarked following three warm-up predictions. The average latency was 6.64 ms, the median latency 6.18 ms, and the P95 latency 7.96 ms, which equates to about 150.50 images per second. The trained checkpoint took up approximately 19.92 MB. The small difference between the median and the P95 latency shows that the GPU inference at the model level was stable under the benchmark conditions and thus provides a good amount of computational headroom for frame-level classification. End-to-end CCTV latency will also involve the acquisition of frames, decoding, resizing, temporal aggregation, network communication, the generation of alerts, logging, and human verification.

5. Discussion

The main experimental result is that a carefully curated, balanced training set can support extremely strong image-level activity classification. The model achieved 99.74% accuracy on an artificially unbalanced test set, while balanced accuracy remained 99.12% and macro F1 99.40%. This matters because it shows performance was not simply achieved by predicting the majority normal class. All four of the hypotheses were supported. Support for H1 came from the macro F1-score of 0.9940. For H2, support was found in the recall rate of 0.9595 on the naturally imbalanced test set. H3 was supported by a mean model latency of 6.64 ms, which is under the 40 ms benchmark threshold. H4 was supported since raising the acceptance threshold to 0.95 caused the accuracy of the accepted samples to rise from 99.737% to 99.824%. These are the pre-determined benchmark criteria, not assertions regarding effectiveness in real-world settings; further validation using the REG CCTV data is still required. Training-only balancing was also of methodological importance since the original training imbalance ratio of 13.751 was lowered to 1.000, with the greatest increase given to suspicious activity. It should be noted that oversampling does not generate any new independent real-world events; rather, it increases the amount of exposure to the existing minority cases. The good results thus show how effective the learning

process is on the curated benchmark, not that the model has seen the full range of suspicious activity which will actually occur in practice. The class that is suspicious should be looked at in more detail. Although it has 100% precision, so that no image from a different class was labeled as suspicious, its recall of 95.95% shows that three of the 74 suspicious images were missed. These three images account for half of the six test errors, and thus the overall accuracy masks a serious and concentrated risk in practice. Two of them were classified as normal and one as vendors, exactly the results most likely to result in a warning being suppressed or downgraded. Suspicious activity is inherently hard to detect since it is partly defined by context and intent rather than by a specific visible action; for example, standing, carrying an object, approaching an enclosure, or interacting with equipment may be entirely legitimate or potentially harmful depending on authorization, location, duration, and the sequence of events. The small number of images in the suspicious test set also makes its recall less reliable than the estimate based on the majority class. Therefore, the results support the need for temporal aggregation, the use of site-specific context, and physical-sensor verification before the approach is used in the field. The confidence analysis offers another important limitation. Many of the errors had a confidence level very close to 1.00, such as all three of the suspicious cases that were missed (0.99997, 0.99992, and 0.99985). The model was therefore not just uncertain when it made its most serious mistakes; it was actually confident in being wrong. This behavior is in line with known issues concerning the calibration of neural networks (Guo et al., 2017). Simply setting a confidence threshold cannot ensure safety, since the 0.95 rule rejected only 0.351% of the test cases and therefore did not eliminate these false negatives. For operational use, alerting must be based on calibration with independent field data, include checks for temporal consistency, incorporate out-of-distribution detection, obtain corroboration from sensors, and involve human verification. The six errors also contribute to temporal modeling. It is not possible for a single image to reliably tell the difference between a person who is simply standing near a protected structure and one who approaches, handles an enclosure, cuts a component, and then leaves with some material. Temporal aggregation, tracking, or an action-recognition model is able to transform repeated frame-level evidence into a decision at the event level. This is in line with current research in the area of action recognition and surveillance, which is increasingly stressing the importance of temporal context (Heo et al., 2026; Ul Amin et al., 2024). Finally, the very low inference latency suggests that the classifier is suitable for an edge-analytics architecture. The T4 benchmark should not be interpreted as a claim that every edge device will achieve 150 images/s. Rather, it shows the classification model is fast enough that the main engineering constraints in deployment are likely video acquisition, temporal reasoning, sensor integration, networking, storage, and governance. The results of the data audit increase the credibility of the benchmark figure. Eliminating the 10 cross-split duplicate files stops exact image reuse from causing the validation or test performance to be inflated, whereas keeping the training-only duplicates does not result in direct evaluation leakage. Moreover, the audit shows that the normal class includes the duplicate records and also makes up the majority of the original training distribution. Because of these two points, the decision to report balanced accuracy, macro F1, MCC, and kappa rather than just relying on accuracy is reinforced. The results from the oversampling also show exactly what the model has actually learned: for suspicious activity, a training exposure of 13.75 was required, as compared with 1.47 for vendors and no additional oversampling in the case of normal activity. Although the resulting balanced training set of 19,554 images makes the optimization less dependent on frequency, it still does not mean that the benchmark is equivalent to a dataset containing thousands of independent cases of suspicious incidents. This point is important in the context of the REG case study since rare real-world vandalism patterns can differ according to infrastructure type, camera position, weather conditions, clothing, group size, and the sequence of interaction – something that cannot be captured by repeatedly using a small subset of public images. We should interpret the results in light of the scope of the task. The 99.74% top-1

accuracy and the 99.40% macro F1 indicate that the curated six-class benchmark is highly separable for the model and the data in question. Yet these figures should not be taken as proof that a general-purpose YOLO11m classifier will attain the same level of performance when applied to operational REG CCTV. Previous research in the areas of human activity recognition and surveillance has highlighted the importance of temporal context, differences between domains, and the difficulty of capturing rare abnormal events. Accordingly, the current results only extend that body of work at the level of image perception: they show that a fast classifier can provide a solid first-stage representation from which a system incorporating temporal and multimodal elements could be built. Direct numerical comparison with unrelated HAR benchmarks would nevertheless be misleading because action definitions, viewpoints, datasets, class counts, split protocols, and metrics differ. The defensible interpretation is that the proposed procedure produced strong internal discrimination under the stated conditions, while REG field performance remains an external-validation question. The class-level results show that suspicious activity is the most difficult operational category. Its precision was 100.00%, but recall was 95.95%, with errors mainly toward normal and vendors. The classifier can therefore be conservative when assigning suspicious, but a single frame lacks enough context to trigger automatic enforcement. The climbing, cutting, and opening categories are not equally important when it comes to risk. A failure to detect a cutting event will probably cause more physical damage than a failure to carry out a normal or vendor observation, while an opening event could indicate either authorized maintenance or unauthorized access. For this reason, the future operational policy should allocate specific costs according to category rather than considering all errors in classification as being the same. The thresholds can be adjusted in light of the consequences of missed events and false alarms; however, these thresholds should be based on events that have been confirmed in the field and not chosen solely on the basis of the benchmark. The confidence analysis strengthens this conclusion. Correct predictions were generally assigned very high confidence, but some incorrect predictions also received confidence close to one. Consequently, a numerical confidence score is not a trustworthy proxy for the probability that a security incident has occurred. Before confidence can become part of an operational risk score, the system needs calibration, temporal persistence, authorization information, and corroborating sensors. This distinction matters especially for public-sector deployment. A system that automatically escalates every high-confidence frame could create unnecessary interventions and undermine trust in the monitoring system. A better design uses the classifier to prioritize evidence for trained operators. Repeated predictions across consecutive frames, movement toward a restricted asset, an access-control mismatch, or a simultaneous physical sensor event can increase alert priority while preserving human oversight. The results also reinforce the central limitation of image-level HAR. An image can show a person near an electricity asset, but it cannot establish the sequence of approach, access, manipulation, damage, and departure. Temporal modeling should therefore be viewed as the next methodological layer rather than as an optional feature. Tracking can maintain object identity across frames; temporal aggregation can suppress isolated errors; and sequence models can learn patterns that distinguish transient presence from sustained interaction. For REG, this staged approach is preferable to immediately replacing the classifier with a much larger end-to-end model. The present benchmark establishes the spatial perception component and provides a measurable baseline. A subsequent field study can then determine whether temporal models improve recall of confirmed incidents without producing unacceptable false-alarm rates. This creates a controlled research progression from image recognition to event recognition. The strongest limitation of the current evidence is domain shift. The training images came from public datasets rather than REG's operational CCTV network. Real REG cameras may differ in height, focal length, compression, illumination, weather exposure, distance to subjects, background, and infrastructure geometry. Authorized technicians may also wear uniforms or carry equipment not present in the public images. Conversely, real incidents may contain

occlusion, multiple people, poor lighting, or unusual viewpoints. Accordingly, the current model should be considered a benchmark prototype rather than a validated REG security product. Field validation should use location-separated and event-separated splits so that frames from the same incident do not appear in both training and test sets. Evaluation should include confirmed legitimate maintenance, ordinary community activity, vendors, and confirmed security incidents. This would allow the model's apparent benchmark performance to be tested against the actual operating conditions that determine usefulness. The REG case study suggests that the classifier's most valuable output is not a binary "thief/not thief" decision, but structured evidence that can support a coordinated response. The model can flag candidate frames, identify the dominant activity class, attach confidence and time information, and pass the event to a temporal decision module. The approximately 19.92-MB checkpoint and 6.64-ms mean inference latency suggest the visual classifier is compact enough for edge-oriented implementation. However, the T4 benchmark cannot be treated as a field-device performance claim. Sensor and asset information can then determine whether the event deserves escalation. This architecture is consistent with the distinction between visual recognition and operational verification established throughout the study. The proposed workflow also separates responsibilities: REG defines protected assets and operational requirements, a research institution can independently validate performance, and Rwanda National Police can participate in the lawful response pathway where appropriate. This reduces the risk of treating an experimental model as autonomous authority and supports auditability. REG is the case-study institution because the activity taxonomy and response requirements arise from documented electricity-infrastructure vandalism in Rwanda (Rwanda Energy Group, 2020, 2021, 2025). A pilot should begin at a limited number of high-risk sites, collect locally representative day and night video under approved governance, and compare classifier-only alerts with temporally aggregated and sensor-corroborated alerts. The pilot must measure event-level recall, false alarms per camera-hour, time to verification, system latency, robustness to weather and occlusion, and operator workload before making any operational claim. A practical deployment should retain existing CCTV infrastructure and add an intelligent analytics layer. The Figure 11 is our proposed study architecture:

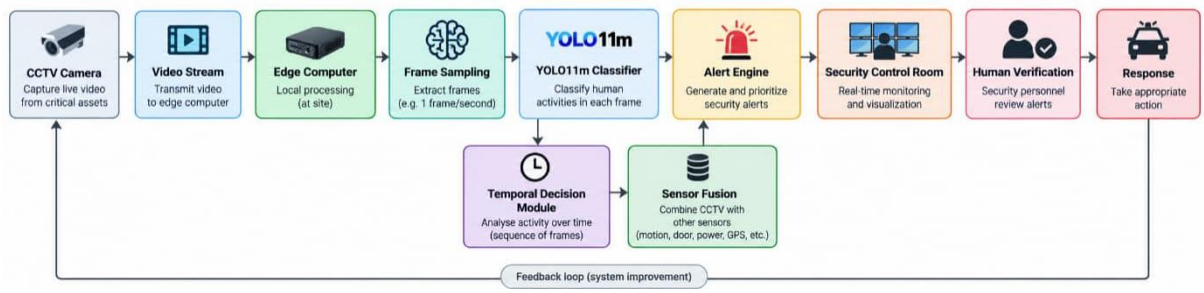

Figure 11: CCTV–AI–sensor architecture

A 25-FPS stream could initially be sampled at 5–10 FPS, with predictions aggregated over a short temporal window. This would reduce redundant computation while preserving enough temporal information to confirm events. Recommended complementary sensors include fence-vibration sensors, magnetic door/contact sensors, electrical measurements, acoustic sensors, PIR/infrared motion detectors, GPS/geofencing, and access-control systems. For example, a suspicious visual prediction combined with fence vibration and an unauthorized access signal should carry greater operational weight than a suspicious prediction alone. A graded policy is recommended: informational status for ordinary or authorized activity, a watch state for repeated or uncertain interaction, and a critical alert only when high-risk visual activity persists or is corroborated by

another signal. Thresholds should be validated on site-specific data, and every alert should preserve the contributing evidence and operator decision for audit.

Privacy, ethics, and governance

CCTV-based AI processing involves personal data and must therefore be governed as carefully as it is engineered. Rwanda's data-protection framework requires lawful, fair, transparent, purpose-specific, proportionate, and secure processing of personal data, with particular attention to large-scale systematic monitoring of publicly accessible areas (Republic of Rwanda, 2021). A deployment should define a legitimate purpose, minimize data collection, establish retention periods, enforce access controls, encrypt sensitive information, maintain audit logs, and implement documented incident-response procedures. The proposed REG system concerns activity recognition, not identity recognition. Facial recognition should not be added merely because cameras make it technically possible. Activity classification can provide an early-warning function without identifying individuals. Any future identity-related functionality should be separately justified, assessed, and governed under the applicable legal and institutional framework. The deployment architecture should therefore be understood as a testable hypothesis about how the benchmark model could be integrated, not as evidence that the complete system has already been validated. The most important future experiment is an instrumented REG pilot in which model alerts can be matched against confirmed events and legitimate activities. Such a pilot would allow real operational evidence to optimize thresholds, temporal windows, sensor-fusion rules, and operator workload.

Limitations and future research

The principal limitation is domain shift: public Kaggle images do not reproduce REG camera height, compression, weather, lighting, infrastructure geometry, authorization patterns, or event prevalence. The incident reports establish context but are not a systematic crime database, and the experiment evaluates isolated frames rather than temporal events. Oversampling repeats minority observations, only one architecture was tested, and latency was measured on a Tesla T4 rather than the intended edge device. The results therefore establish an internal benchmark, not field effectiveness, causal crime reduction, or economic benefit. The next research stage should collect a REG electricity-infrastructure CCTV dataset under controlled and ethically approved conditions, with day/night variation, multiple camera angles and distances, multiple infrastructure types, authorized maintenance, legitimate vendors, and realistic suspicious activities. Split the dataset by location and event where possible, rather than by near-identical frames. Then evaluate temporal models such as LSTM/ConvLSTM, temporal transformers, temporal convolutions, or tracking-based aggregation. Multimodal fusion should combine visual, vibration, electrical, acoustic, PIR, contact, access-control, and environmental evidence. Finally, the model should be benchmarked on practical edge hardware and continuously evaluated using confirmed false positives and false negatives from pilot deployments.

6. Conclusion

This study developed a six-class human activity recognition system intended to serve as a visual early-warning component for electricity-infrastructure protection for Rwanda Energy Group (REG). We curated two public Kaggle image sources into climbing, cutting, normal, opening, suspicious, and vendors classes. After cleaning, the dataset contained 23,647 images: 19,554 balanced training images, 1,814 validation images, and 2,279 test images. Training-only oversampling reduced the imbalance ratio from 13.751 to 1.000 without altering the validation or test distributions. A pretrained YOLO11m classification model trained with 224 × 224 inputs, augmentation, AdamW, dropout, cosine scheduling, and early stopping achieved 99.74% test accuracy, 99.12% balanced accuracy, 99.40% macro F1, 0.9998 macro-ROC-AUC, and

approximately 0.9971 macro average precision. Only six test images were misclassified. Mean inference latency was 6.64 ms on a Tesla T4, with 7.96 ms P95 latency and approximately 150.50 images/s throughput. These results establish a strong image-level perception baseline under the curated benchmark conditions. Within the benchmark scope, the findings support the six empirical research questions and all four hypotheses. The taxonomy was learnable, training-only balancing preserved an untouched evaluation distribution, YOLO11m separated the six classes with very high aggregate performance, and model-level inference was fast. However, the concentrated suspicious-class errors, high-confidence mistakes, and public-dataset domain shift prevent any claim that a complete early-warning or theft-prevention system has been validated. For REG, the practical implication is a staged architecture in which fast frame classification is followed by temporal reasoning, sensor fusion, asset and authorization context, alert prioritization, and human verification. The reported results justify progressing to a controlled field study, but they do not establish operational accuracy on REG CCTV. The next research stage should collect ethically governed, location-separated REG footage containing both confirmed incidents and legitimate activity, and should evaluate temporal and multimodal models against field-confirmed events. This progression would turn the present benchmark contribution into evidence for a deployable REG electricity-infrastructure early-warning system.

7. Declarations

Ethics approval and consent to participate. Not applicable; the study used public image datasets and did not collect new human-subject data.Data availability. The study used the public Wall Climbing ALEE / Wall Security Dataset ALEE and Climbing V2 datasets. The curated dataset itself is not claimed to be a nationally representative Rwandan CCTV dataset.

8. References

AlMuhaideb, S., AlAbdulkarim, L., AlShahrani, D. M., AlDhubaib, H., & AlSadoun, D. E. (2024). Achieving more with less: A lightweight deep learning solution for advanced human activity recognition (HAR). *Sensors, 24(16), 5436.* https://doi.org/10.3390/s24165436

Alomar, K., Aysel, H. I., & Cai, X. (2025). CNNs, RNNs, and Transformers in human action recognition: A survey and a hybrid model. *Artificial Intelligence Review, 58, 387.* https://doi.org/10.1007/s10462-025-11388-3

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. Interacting with Computers, 23(1), 4–17. https://doi.org/10.1016/j.intcom.2010.07.003

Buningwire, W. (2019). Rwanda Energy Group announces RWF 19bn power loss. *KT Press.* https://www.ktpress.rw/2019/01/rwanda-energy-group-announces-rwf-19bn-power-loss/

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the Kinetics dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4724–4733).* https://doi.org/10.1109/CVPR.2017.502

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9260–9269).* https://doi.org/10.1109/CVPR.2019.00949

Duja, K. U., Khan, I. A., & Alsuhaibani, M. (2024). Video surveillance anomaly detection: A review on deep learning benchmarks. *IEEE Access, 12, 164811–164842.* https://doi.org/10.1109/ACCESS.2024.3491868

Dutta, S. J., Boongoen, T., & Zwiggelaar, R. (2025). Human activity recognition: A review of deep learning-based methods. *IET Computer Vision, 19(1), e70003.* https://doi.org/10.1049/cvi2.70003

El-Adawi, E., Essa, E., Handosa, M., & Elmougy, S. (2024). Wireless body area sensor networks-based human activity recognition using deep learning. *Scientific Reports, 14, 2702.* https://doi.org/10.1038/s41598-024-53069-1

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6201–6210).* https://doi.org/10.1109/ICCV.2019.00630

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning, 70, 1321–1330.* https://doi.org/10.48550/arXiv.1706.04599

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).* https://doi.org/10.1109/CVPR.2016.90

Heo, S., Moon, J., & Jung, S. K. (2026). Action recognition: A comprehensive survey of tasks, methods, and challenges. *ICT Express, 12(1), 32–49.* https://doi.org/10.1016/j.icte.2025.11.015

Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications, 244, 122778.* https://doi.org/10.1016/j.eswa.2023.122778

Kumar, T., Brennan, R., Mileo, A., & Bendechache, M. (2024). Image data augmentation approaches: A comprehensive survey and future directions. *IEEE Access, 12, 187536–187571.* https://doi.org/10.1109/ACCESS.2024.3470122

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision (pp. 2980–2988). https://doi.org/10.1109/ICCV.2017.324

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video Swin Transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3202–3211).* https://doi.org/10.1109/CVPR52688.2022.00320

Mumuni, A., Mumuni, F., & Gerrar, N. K. (2024). A survey of synthetic data augmentation methods in machine vision. *Machine Intelligence Research, 21(5), 831–869.* https://doi.org/10.1007/s11633-022-1411-7

ngcduy09on. (2024). Climbing V2 [Data set]. *Kaggle.* https://www.kaggle.com/datasets/ngcduy09on/climbing-v2

Nguyen, H.-C., Nguyen, T.-H., Scherer, R., & Le, V.-H. (2023). Deep learning for human activity recognition on 3D human skeleton: Survey and comparative study. *Sensors, 23(11), 5121.* https://doi.org/10.3390/s23115121

Nissenbaum, H. (2004). Privacy as contextual integrity. Washington Law Review, 79(1), 119–158.

Qureshi, A. N. (2026). Wall Climbing ALEE / Wall Security Dataset ALEE [Data set]. *Kaggle.* https://www.kaggle.com/datasets/alinoorqureshi/wall-climbing-alee

Republic of Rwanda. (2021). Law No. 058/2021 relating to the protection of personal data and privacy. *Rwandan legislation.* https://www.risa.gov.rw/data-protection-and-privacy-law

Rwanda Energy Group. (2020). Rwanda Energy Group (REG) condemns theft of electricity as it causes both revenue losses and fatal accidents. *REG News.* https://www.reg.rw/media-center/news-details/news/rwanda-energy-group-reg-condemns-theft-of-electricity-as-it-causes-both-revenue-losses-and-fatal-a/

Rwanda Energy Group. (2021). Four people arrested in Kigali over vandalizing and stealing electricity pylons. *REG News.* https://www.reg.rw/media-center/news-details/news/four-people-arrested-in-kigali-over-vandalizing-and-stealing-electricity-pylons/

Rwanda Energy Group. (2025). REG warns: Electricity vandalism threatens economy and essential services delivery. *REG News.* https://www.reg.rw/media-center/news-details/news/reg-warns-electricity-vandalism-threatens-economy-and-essential-services-delivery/

Rwanda National Police. (2020). Three arrested for alleged vandalism of public infrastructure. *Rwanda National Police.* https://police.gov.rw/media/news-detail/news/three-arrested-for-alleged-vandalism-of-public-infrastructure/

Rwanda National Police. (2021a). RNP media weekly review. *Rwanda National Police.* https://police.gov.rw/media/news-detail/news/rnp-media-weekly-review-b753dfe357/

Rwanda National Police. (2021b). Rusizi: Man arrested for vandalizing fiber-optic cables. *Rwanda National Police.* https://police.gov.rw/media/news-detail/news/rusizi-man-arrested-for-vandalizing-fiber-optic-cables/

Rwanda National Police. (2022). Kigali: Trio arrested over electricity infrastructure vandalism. *Rwanda National Police.* https://police.gov.rw/media/news-detail/news/kigali-trio-arrested-over-electricity-infrastructure-vandalism/

Rwanda National Police. (2023a). Kicukiro: Five arrested for vandalizing electricity infrastructure. *Rwanda National Police.* https://police.gov.rw/media/news-detail/news/kicukiro-five-arrested-for-vandalizing-electricity-infrastructure/

Rwanda National Police. (2023b). Security in Rwanda continues to improve - Minister Gasana. *Rwanda National Police.* https://police.gov.rw/media/news-detail/news/security-in-rwanda-continues-to-improve-minister-gasana/

Rwanda National Police. (2024). RNP, City of Kigali meet scrap dealers to fight vandalism of public infrastructure. *Rwanda National Police.* https://police.gov.rw/media/news-detail/news/rnpcityofkigalimeetscrapdealerstofightvandalismofpublicinfrastructure/

Sedaghati, N., Ardebili, S., & Ghaffari, A. (2025). Application of human activity/action recognition: A review. *Multimedia Tools and Applications, 84, 33475–33504.* https://doi.org/10.1007/s11042-024-20576-2

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6479–6488).* https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00678

Surek, G. A. S., Seman, L. O., Stefenon, S. F., Mariani, V. C., & Coelho, L. D. S. (2023). Video-based human activity recognition using deep learning approaches. *Sensors, 23(14), 6384.* https://doi.org/10.3390/s23146384

Ul Amin, S., Kim, B., Jung, Y., Seo, S., & Park, S. (2024). Video anomaly detection utilizing efficient spatiotemporal feature fusion with 3D convolutions and long short-term memory modules. *Advanced Intelligent Systems, 6(7), 2300706.* https://doi.org/10.1002/aisy.202300706

Ultralytics. (2025). YOLO11 documentation and classification models. *Ultralytics.* https://docs.ultralytics.com/models/yolo11/

Varga, D. (2025). Decision fusion-based deep learning for channel state information channel-aware human action recognition. *Sensors, 25(4), 1061.* https://doi.org/10.3390/s25041061

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7794–7803).* https://doi.org/10.1109/CVPR.2018.00813

Wastupranata, L. M., Kong, S. G., & Wang, L. (2024). Deep learning for abnormal human behavior detection in surveillance videos—A survey. *Electronics, 13(13), 2579.* https://doi.org/10.3390/electronics13132579

Wazwaz, A., Amin, K., Semary, N., & Ghanem, T. (2024). Dynamic and distributed intelligence over smart devices, Internet of Things edges, and cloud computing for human activity recognition using wearable sensors. *Journal of Sensor and Actuator Networks, 13(1), 5.* https://doi.org/10.3390/jsan13010005

Zeng, W. (2024). Image data augmentation techniques based on deep learning: A survey. *Mathematical Biosciences and Engineering, 21(6), 6190–6224.* https://doi.org/10.3934/mbe.2024272

Published by

IJSAB International