A Comparative Study of Vision Transformers and Multimodal Language Models for Violence Detection in Videos

Tomas Ditchfield-Ogle
School of Computing and Communications
Lancaster University
Lancaster, United Kingdom
tomas.o.ogle@gmail.com

Ruslan Mitkov
University of Alicante
ruslan.mitkov@ua.es

Abstract

This study compares methods for detecting violent videos, which are crucial for ensuring real-time safety in surveillance and digital moderation. It evaluates four approaches: a random forest classifier, a transformer model, and two multimodal vision-language models. The process involves preprocessing datasets, training models, and assessing accuracy, interpretability, scalability, and real-time suitability. Results show that traditional methods are simple but less effective. The transformer model achieved high accuracy, and the multimodal models offered high violence recall with descriptive justifications. The study highlights trade-offs and provides practical insights for the deployment of automated violence detection.

1 Introduction

Concerns about harmful content have prompted the UK government to implement the Online Safety Act 2023, which encourages proactive content moderation and violence prevention both online and offline (GOV UK Department for Science, 2025). As smart cities evolve, citizens demand enhanced safety measures and swift emergency responses, pressuring authorities to adopt automation tools (Pujol et al., 2020). Governments are facing the rapid growth of video content in surveillance and digital applications, making manual analysis impractical. This drives the need for real-time video systems that identify patterns for safety and emergencies (Sabha and Selwal, 2024). Social media platforms also struggle to manage vast video volumes in near real-time (Pujol et al., 2020), amid increasing circulation of violent content, including hate crimes and terrorist attacks (Studer, 2017). Many platforms, such as Facebook and YouTube, attempt to moderate content through automated tools; however, the scale and immediacy of live streaming make this nearly impossible (Pujol et al., 2020).

Automatic violence detection is difficult due to its inherent subjectivity. Violent acts are not always visually explicit and depend on context, like body posture, group dynamics, or weapons, posing barriers to definition (Naik and Gopalakrishna, 2017). Other issues include illumination variance, which affects outdoor video quality due to changes in lighting, such as day/night transitions or weather, impacting colour and contrast (Kaur and Singh, 2024).

Fortunately, AI offers promising tools, particularly through computer vision (CV) and machine learning models trained to classify visual data. This project explores and compares four such methods for detecting violence: Random Forest classifier, TimeSformer, Llama 3.2

Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models associated with RANLP 2025, pages 10–20, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-102-8-002

Vision Instruct and Janus Pro. This project aims to investigate the effectiveness of cutting-edge machine learning technologies in detecting real-world violence. This research has applications ranging from enhancing online moderation to improving street safety in cities.

2 Background

Recent research indicates that machine learning models are increasingly supporting video classification, particularly in the context of violence detection (VD).

2.1 Violence detection (VD) software

Before Deep Learning (DL) methods, VD was seen as recognising specific human actions (Peixoto et al., 2020). Following initial approaches, DL techniques improved VD results (Peixoto et al., 2020), notably with 3D convolutional neural networks for extracting spatio-temporal patterns (Ding et al., 2014). However, many models remain computationally intensive, often using multi-stream input and stacked LSTM layers, with limited details on their complexity (Ullah et al., 2019) (Ullah et al., 2022). Some researchers focus on models balancing performance and efficiency. One achieved 87.25% accuracy with just 0.27 million parameters (Cheng et al., 2021a) using depth-wise separable convolutions from Pseudo-3D Residual Networks (Qiu et al., 2017) and MobileNet (Howard et al., 2017), thereby reducing complexity without compromising accuracy. The VD field has evolved from handcrafted features to advanced DL models that interpret video spatio-temporal cues.

2.2 Random Forest

Random Forests are a popular machine learning model used for classification and forecasting, requiring high-quality data for training. They improve algorithms and user behaviour analysis, aiding pattern recognition (Salman et al., 2024). The model excels in classification and regression, using cross-validation for accuracy and handling missing data effectively (Achari and Sugumar, 2024). It also reduces bias by training multiple decision trees on random subsets of data, making it one of the most reliable techniques in ML (Salman et al., 2024). Random forests are often used in hybrid models for VD. For example, a study developed a facial recognition assault system using Random Forest, achieving 98% precision and 97.5% accuracy, showing ensemble methods enhance safety (Ohwosoro et al., 2024).

2.3 TimeSformer

Video understanding tasks, such as VD, require models to interpret spatial and temporal features. TimeSformers, which utilise a transformer-based architecture with spatio-temporal attention, reason across frames and time (Bertasius et al., 2021a). Research has found TimeSformer performs well in DeepFake detection (Chen et al., 2024). The architecture is suitable for VD, where an extended temporal context is key. TimeSformer differs from standard Transformers in that it learns spatio-temporal features directly from frame patches. Research shows that "divided attention," applying temporal and spatial attention separately, achieves the highest video classification accuracy (Bertasius et al., 2021b).

2.4 Large Language Models

Following ChatGPT's launch, attention has focused on large language models (LLMs) (Tian et al., 2024), especially for their strong performance in classification (Al Faraby et al., 2024), summarisation (Doss et al., 2024), data and code generation (Shimabucoro et al., 2024) (Nejjar et al., 2025). The rapid development of LLMS is clear in the late 2023 and early 2024 releases of Google's Gemini, Anthropic's

Claude 3, and OpenAI's GPT-4 (Shahriar et al., 2024). These models represent a significant leap in capabilities, transitioning from text-only to multimodal understanding across text, images, and audio, with enhanced parameters and speed (Shahriar et al., 2024). LLMs' ability to understand and generate extensive data has created opportunities, such as Llama3.2, which addresses predatory conversations and abusive messages (Arora, 2025). However, these models can gain vision capabilities. Vision in LLMs means adapting transformer models from language to interpret images (Yenduri et al., 2024). This has expanded the generative pre-trained transformer (GPT) to include vision. Since multimodal LLMs are relatively new, many research areas are still in their early stages (Wang et al., 2024). OpenAI's GPT-4 release in May 2024 marked a key shift, as it was the first to interpret emotions from videos (Islam and Moushi, 2024), opening up new applications. Yet, using multimodal LLMs for VD in videos remains under-explored, with research gaps this project aims to fill (Jaafar and Lachiri, 2023).

3 Data

The violent samples in both datasets depict real-world street fight scenarios recorded under varying conditions. Non-violent samples include everyday activities like walking, eating, and playing sports, representing a wide range of non-aggressive behaviours. This diversity provides a realistic setting for evaluating safety monitoring and automated incident detection systems.

3.1 Ethical concerns

All datasets used in this study were obtained from publicly available academic sources. No new data was collected, annotated, or shared during the project. The RLVS dataset was sourced from Kaggle (Mustafa, 2020) and introduced initially by Soliman et al. (Soliman et al., 2019). The RWF-2000 dataset was downloaded from Hugging Face and first presented by Cheng et al. (Cheng et al., 2021b). Both datasets contain publicly available videos designed for violence detection research. None of the content includes personally identifiable information, as all videos were either anonymised or publicly accessible.

3.2 RLVS Dataset

A subset of the Real-Life Violence Situations (RLVS) dataset comprising 957 violent samples and 839 non-violent samples was used for training, validation, and in-distribution testing. As a result, the final dataset used in this study. A consistent train/validation/test split was generated and saved in a persistent JSON file to support reproducibility.

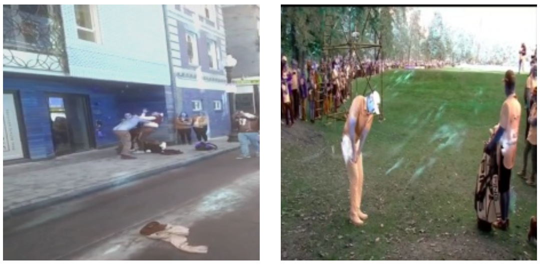


Figure 1: Example frames from RLVS. Left: violent, Right: non-violent

3.3 RWF-2000 Dataset

The RWF-2000 dataset was used solely for out-of-distribution testing to assess generalisation beyond the training data. A subset of 383 violent and 395 non-violent was used for testing to reduce computational load, especially for the vision-language models. Models were neither trained nor fine-tuned on RWF-2000, and no manual labelling or editing was done.

4 Methodology

4.1 Data Preprocessing

A unified pipeline ensured consistent inputs across models. Videos were uniformly sampled every 15 frames, with up to 16 frames per clip. Frames were resized to 224 × 224; clips with fewer than 8 valid frames were discarded. Training and validation sets were augmented with brightness, contrast, and saturation shifts (±20%), hue shifts (±0.1), horizontal flips, random crops, and rotations (±10°), applied consistently across frames to preserve temporal coherence. RLVS was split via stratified sampling (20% test, 10% validation); for RWF-2000, a subset was used for testing only. Motion features were derived from Farneback optical flow, summarised with statistics (mean, variance, skewness, range, etc.) and a high-motion pixel count. Frames were saved as JPEGs (for LLMs) and as tensors ($T \times C \times H \times W$) in PyTorch format for efficient loading.

4.2 Random Forest

To establish a classical baseline, a Random Forest classifier was trained on motion features derived from dense optical flow.

4.2.1 Training

The Random Forest model was trained on motion statistics derived from dense optical flow, including mean, median, standard deviation, maximum, minimum, range, skewness, variance, and the proportion of high-motion pixels per frame. These features capture both overall motion intensity and its distribution across frames. Labels were assigned automatically from dataset filename prefixes ('violent-' or 'nonviolent-'), consistent with the dataset's original annotation scheme. Hyperparameters were optimised via grid search with five-fold cross-validation, using ROC-AUC as the scoring metric due to class imbalance. The best model employed 100 estimators, a maximum depth of 8, a minimum sample split of 10, a minimum sample leaf of 4, and balanced class weighting. This configuration was retrained on the whole training set and evaluated on the test set.

4.2.2 Feature Importance

Post-training, feature importances revealed that mean and minimum motion magnitudes were the most influential predictors, underscoring the role of motion intensity in distinguishing between violent and non-violent activity.

4.2.3 Interpretability of Random Forest

The Random Forest provides insight into which features drive classification. For example, a high mean motion magnitude strongly predicted violent sequences, such as street fights, whereas a low minimum flow magnitude aligned with stable, non-violent scenes. However, the model also produced false positives in contexts like crowd surges at sports events, where collective movement mimicked aggression. These results suggest that RF's interpretability is valuable, but its rule-like motion thresholds are not robust across diverse scenarios.

4.3 TimeSformer

To establish a DL benchmark, a transformer-based video classification model was implemented using the TimeSformer architecture. TimeSformer builds on the Vision Transformer (ViT) framework by introducing a mechanism to handle both spatial and temporal dimensions in video data. Rather than using traditional 3D convolutions, it applies divided attention across space and time separately, enabling efficient and scalable video understanding from raw pixel data.

4.3.1 Model Configuration

The TimeSformer model used was `facebook/TimeSformer-base-finetuned-k400`, pre-trained on the Kinetics-400 dataset (Bertasius et al., 2021a). To adapt it for binary violence detection, the classification head was replaced with a fully connected layer producing two logits. Frames were converted to floating point and normalised by dividing by 255.0, preserving dynamic range without distorting pixel intensities. Temporal tensors were zero-padded as needed. Labels were inferred from filename prefixes. The entire model was fine-tuned to adapt specifically to the task.

4.3.2 Training Configuration

Training used the Hugging Face Trainer API with a batch size of 6. The model was optimised for cross-entropy loss with label smoothing (0.1), a learning rate of 5e-5, and cosine scheduling, along with a 25% warm-up. Early stopping had a patience of 2 steps. Regularisation included weight decay (0.2) and gradient clipping (norm 1.0). Evaluation occurred every 1000 steps, saving the best model based on validation log loss. Seeds were fixed at 42 for reproducibility. Training ran on a SLURM job with an NVIDIA A5000 GPU.

4.4 Interpreting TimeSformer decisions

We analyse the model's posterior $p(\text{violence} \mid x)$ without binarisation. To expose its decision process, we extract self-attention from each transformer block during a forward pass. (`output_attentions=true`) and apply attention rollout: heads are averaged, an identity residual added, rows normalised, and attention matrices multiplied across layers to form a single CLS→patch relevance map. This is reshaped into a $g \times g$ grid and temporal tubelets to yield spatial heatmaps (time-averaged). In the Kinetics-400 TimeSformer, inputs are 224×224 with 16×16 patches ($g = 14$) and tubelet size 8, so 16 frames give $T_{\text{eff}} = 2$ temporal tokens inferred at run time. For windowed videos, video-level probabilities are aggregated with monotone poolers that preserve probabilistic semantics: $\max$, mean, top-k mean, log-sum-exp (temperature τ), and noisy-OR $1 - \prod_i (1 - p_i)$. Evaluation utilities proper scoring rules (negative log-likelihood, Brier) and calibration/ranking metrics (ROC–AUC, PR–AUC, ECE/MCE).

4.5 Multimodal LLMs: Llama 3.2 Vision and Janus-Pro-7B

To assess the potential of multimodal large language models for violence detection, Meta's Llama 3.2 Vision Instruct model and DeepSeek's Janus-Pro-7B were used.

4.5.1 Llama 3.2 Vision Instruct Model Configuration

The model (`meta-Llama/Llama-3.2-11B-Vision-Instruct`) was loaded via Hugging Face Transformers (meta, 2024) with mixed-precision evaluation. Inputs were processed using AutoProcessor for image normalisation and prompt tokenisation. Generation was limited to 200 tokens with deterministic decoding (do_sample=False, temperature=0.2) to ensure stable outputs. Each frame was evaluated using the following prompt: ”This image is part of a public dataset of street and public scenes used for academic research. Start your response with a yes or no if violence is depicted in this image. Then describe what is happening. If a violent or aggressive incident occurs, describe what happened and identify those involved. If there isn't any violence, describe the scene as peaceful or non-violent. Use simple language and avoid complex terms where possible.”

4.5.2 Janus-Pro-7B Vision Model Configuration

Janus was loaded using AutoModelForCausalLM with mixed precision enabled. The Janus-specific VLChatProcessor was used to process images and chat-style prompts, ensuring consistent resizing, normalisation, and tokenisation. The prompt used was identical to that used with Llama. Generation parameters were configured with do_sample=False, repetition_penalty=1.0, and a maximum of 200 new tokens to produce deterministic and focused outputs.

4.6 Testing Methodology

To evaluate model effectiveness and generalisability, two testing settings were used: in-domain testing on the RLVS test split and out-of-domain testing on a subset of RWF-2000. This allowed assessment of performance within the original data distribution and in unseen environments. A unified preprocessing and evaluation pipeline standardised video extraction, transformation, and organisation for both datasets. The RLVS test set consisted of 363 videos (194 violent and 169 non-violent), which were held out from training and validation. The RWF-2000 subset included 778 videos (383 violent, 395 non-violent), enabling fair cross-model comparison. To improve robustness and simulate real-world variability, data augmentation was applied at the video level with a 50% probability during RLVS training and validation. Extracted frames were then formatted as inputs for the three models. For input preparation, the Random Forest model used optical flow between consecutive frames to extract nine motion statistics forming fixed-length feature vectors. The TimeSformer model received RGB frame tensors of shape (16, 3, 224, 224) and applied spatial and temporal self-attention for classification. Llama and Janus processed frames individually with a fixed prompt; a zero-shot text classifier classified their generated text outputs to assign violent or non-violent labels. Outputs were compared to ground truth labels, with confusion matrices used to analyse false positives and negatives. Performance metrics included accuracy, precision, recall, and F1-score.

5 Evaluation

5.1 In-Domain Testing (RLVS)

Model	Accuracy	Precision	Recall	F1	Time (s)
Random Forest	77.96%	0.7614	0.8556	0.8058	0.01
TimeSformer	96.41%	0.9547	0.9793	0.9669	39.90
LLaMA	78.24%	0.7154	0.9845	0.8286	99 309.80
Janus Pro	74.38%	0.6772	0.9948	0.8058	9 904.08

Table 1: In-domain RLVS test set performance.

Table 1 summarises model performance on the RLVS test set. TimeSformer performed strongest, achieving high accuracy and a balanced precision–recall trade-off with inference times suitable for near-real-time surveillance. LLaMA and Janus Pro reached very high recall, but this came at the cost of precision, often misclassifying non-violent group behaviour as violent. Random Forest was the fastest model, classifying samples almost instantly; however, its reliance on simple motion statistics made it prone to errors in ambiguous scenarios, such as crowd surges. These results suggest that TimeSformer is best suited for automated monitoring, while multimodal models may be more valuable in forensic review or moderation contexts where interpretability is prioritised. Random Forest, despite weaker performance, remains attractive for highly constrained deployments. The error distributions for each RLVS model are illustrated in the corresponding confusion matrices (Figure 2), which make explicit the balance between false positives and false negatives discussed above.

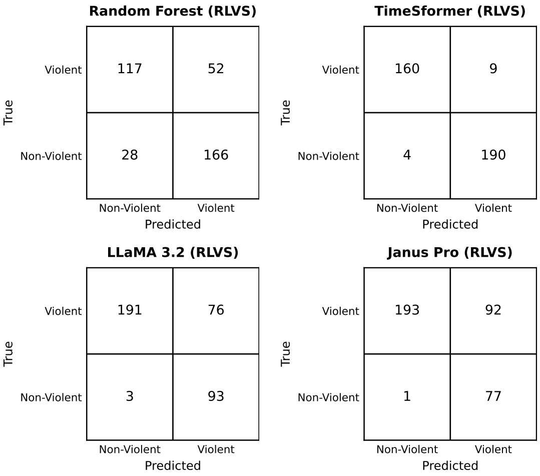


Figure 2: Confusion matrices on the RLVS test set (rows = true labels, columns = predicted labels).

5.2 Out-Of-Domain Testing (RWF-2000)

Model	Accuracy	Precision	Recall	F1	Time (s)
Random Forest	54.24%	0.5754	0.2689	0.3665	0.01
TimeSformer	68.76%	0.6590	0.7571	0.7047	82.16
LLaMA	64.78%	0.5873	0.9635	0.7298	193 729
Janus Pro	74.68%	0.6691	0.9635	0.7898	21 706

Table 2: Out-of-domain RWF-2000 test set performance.

Table 2 presents the performance of all models on the RWF-2000 dataset. Janus Pro achieved the highest F1 score (0.79) with near-perfect recall (0.96), demonstrating strong zero-shot transfer capabilities. LLaMA achieved similar recall but with lower precision, resulting in a higher number of false positives. Both models generated interpretable outputs, though their runtimes were extremely high. TimeSformer generalised well, despite being fine-tuned only on RLVS, achieving balanced scores and completing inference in just over a minute. Random Forest performed poorly under distribution shift, with low recall and F1, reflecting limited robustness. Overall, Janus Pro showed the strongest zero-shot generalisation, while TimeSformer offered a better balance of speed and accuracy. LLaMA remained interpretable, but it was computationally intensive. Random Forest remained the most efficient but least adaptable. Error patterns under distribution shift are shown in the RWF-2000 confusion matrices (Figure 3), which emphasise the models' differing capacities to generalise.

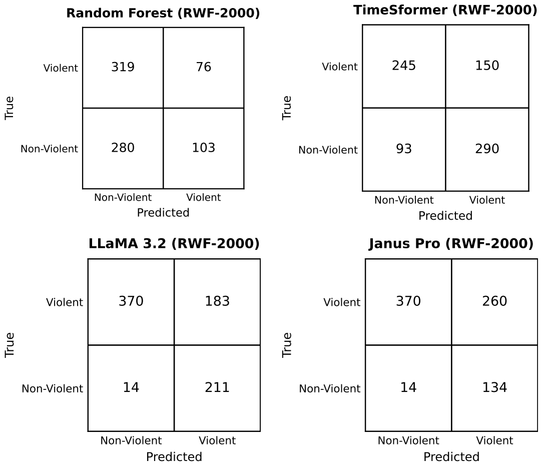


Figure 3: Confusion matrices on the RWF-2000 test set (rows = true labels, columns = predicted labels).

5.3 Decision evidence and probability quality

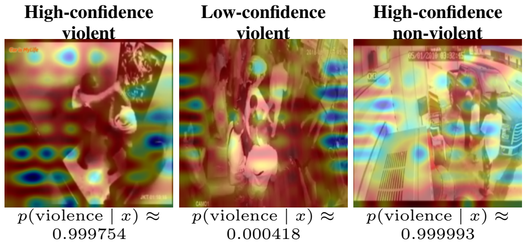
On 778 test windows (383 violent; 395 non-violent), the model yields ROC–AUC 0.767 and PR–AUC 0.764 from raw posteriors. Probability quality is moderate (negative log-likelihood 1.94; Brier 0.288) and calibration indicates over-confidence (ECE 0.286, 15 bins). Cumulative gains show that the top 10% of windows by $p(\text{violence} \mid x)$ contain 18.3% of violent windows (Lift@10% 1.83). Aggregating windows improves video-level ranking: *noisy OR* reaches ROC–AUC 0.800 (PR–AUC 0.751), while *top-k mean* ($k = 3$) gives the best proper scoring (negative log-likelihood 1.851; Brier 0.287) and the lowest ECE among the poolers.

Pooler	ROC–AUC ↑	NLL ↓
max	0.787	2.011
mean	0.755	1.854
noisy–OR	**0.800**	2.149
log–sum–exp	0.759	1.864
top-k mean	0.761	**1.851**

Table 3: Video-level pooling of window probabilities (no thresholds). Best per column in bold.

5.4 Qualitative evidence

Spatial overlays for high-confidence violent windows focus on converging bodies and limbs, with temporal peaks in the tubelet that captures contact. Low confidence violent windows show diffuse attention, often in pre- or post-event frames, under occlusion, or when brief actions are split across tubelets. High confidence non-violent windows emphasise crowd surges or celebratory gestures that are visually salient yet non-violent. Figure 4 shows examples.



High-confidence violent **Low-confidence violent** **High-confidence non-violent**

$p(\text{violence} \mid x) \approx 0.999754$ $p(\text{violence} \mid x) \approx 0.000418$ $p(\text{violence} \mid x) \approx 0.999993$

Figure 4: Spatial attention overlays (CLS→patch relevance) for three representative windows.

5.5 Evaluation Findings

5.5.1 Notable Observations

TimeSformer consistently achieved the best balance of precision and recall across both datasets. Its confusion matrices indicated lower rates of false positives and false negatives, supporting its reliability in varied scenarios. LLaMA Vision exhibited a strong bias towards recall, detecting violent content aggressively but occasionally misclassifying benign scenes as violent. This trade-off may be acceptable in high-sensitivity contexts but is less suitable where false positives carry a significant cost. Random Forest performed reliably on RLVS, particularly in identifying non-violent scenes, but its accuracy declined on RWF-2000. This shows its handcrafted features generalise poorly to more varied or noisy data.

5.5.2 Error Analysis and Interpretability

TimeSformer's errors primarily resulted from crowded or celebratory scenes, which produced false positives, and short, low-contrast violent clips, which resulted in false negatives. LLaMA and Janus Pro often hallucinated aggression, labelling cricket games as violent. Random Forest struggled with camera shake and noisy backgrounds, exposing its reliance on clean motion signals. Interpretability also varied. TimeSformer's attention maps highlighted human interactions, usually aligning with the source of violence. LLaMA and Janus Pro generated natural language explanations, offering detailed scene descriptions of actors, actions, and context, such as environments and expressions. These outputs exposed systematic biases and helped diagnose false positives. They also added value in human-in-the-loop scenarios where moderators could review justifications alongside predictions.

5.5.3 Summary of Findings

TimeSformer delivered the highest overall performance, with strong generalisation and the fewest false positives. Its ability to model spatial and temporal features makes it well-suited for continuous, high-precision surveillance in environments where alert reliability is critical.

LLaMA Vision and Janus Pro achieved the highest recall, demonstrating strong sensitivity to violent content and producing interpretable natural language explanations. These qualities make them valuable for content moderation and investigative or regulatory settings, where comprehensive flagging and explanation are prioritised. However, their lower precision and very high inference times limit their suitability for real-time or autonomous applications.

Random Forest, while fast and transparent, generalised poorly to RWF-2000. Its simplicity and efficiency still make it viable for controlled edge deployments, such as low-power CCTV units, where latency and interpretability take precedence over accuracy. Overall, these findings emphasise distinct deployment niches: TimeSformer as the most balanced and scalable solution, multimodal LLMs for human-in-the-loop systems, and Random Forest for resource-constrained contexts. Together, they illustrate the trade-offs between accuracy, interpretability, and efficiency that must guide real-world adoption of violence detection systems.

6 Conclusion

6.1 Project Limitations

This study has several limitations. The models were not trained to detect weapons, as this was not included in the datasets used in this project, which limits their ability to detect armed violence. No post-hoc calibration was applied, ensuring fairness across models but potentially constraining accuracy and generalisability. Attention maps serve as explanatory aids rather than causal attributions but consistently emphasise physical interaction.

6.2 Future Work

Future work should evaluate these models in real-time surveillance or moderation settings. Adding audio cues, such as raised voices, could support earlier detection. Another approach is to incorporate textual commentary from speech transcripts or subtitles, as verbal threats often precede violence. LLMs can process text and video jointly, enabling cross-modal reasoning. In contrast, models like TimeSformer would need auxiliary NLP components or architectural changes. Methodological steps include aligning subtitles with video frames, fine-tuning multimodal encoders, and comparing late-fusion against joint-embedding approaches to determine which best captures temporal and semantic dependencies. Such integration could provide richer context and improve robustness in safety-critical applications.

6.3 Summary

This project compared four approaches to AVD in video: a Random Forest baseline, the transformer-based TimeSformer, LLaMA 3.2 Vision Instruct and Janus Pro, evaluated on RLVS and RWF-2000 datasets. TimeSformer achieved the strongest balance of accuracy and efficiency, making it suitable for real-world deployment. LLaMA Vision demonstrated high recall and interpretability, which is valuable in settings with human oversight; however, computational demands limit its scalability. The Random Forest was lightweight and interpretable but struggled to generalise, highlighting the limits of handcrafted features. Overall, transformer-based models appear most promising when balancing performance and scalability. Future directions include model distillation, real-time optimisation, and audio integration.

Acknowledgements

This work has been partially supported by the CIDEXG/2023/12 project, funded by the Generalitat Valenciana. We thank Dexter Latcham and Suzannah Read for their reading of earlier drafts and editorial suggestions; any remaining errors are our own.

References

A Prudhvi Sai Kumar Achari and R Sugumar. 2024. Performance analysis and determination of accuracy using machine learning techniques for naive bayes and random forest. In *AIP Conference Proceedings*, volume 3193, page 020199. AIP Publishing LLC.

Said Al Faraby, Ade Romadhony, et al. 2024. Analysis of llms for educational question classification and generation. *Computers and Education: Artificial Intelligence*, 7:100298.

Ankush Arora. 2025. *Detecting Online Abuse: Fine-Tuning LLMs for Abusive Language Detection*. Ph.D. thesis, Universität Koblenz.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021a. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, pages 813–824. PMLR.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021b. Is space-time attention all you need for video understanding?

Zhengxuan Chen, Shuo Wang, Deyang Yan, and Yushi Li. 2024. A spatio-temporl deepfake video detection method based on timesformer-cnn. In *2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, pages 1–6. IEEE.

Ming Cheng, Kunjing Cai, and Ming Li. 2021a. Rwf-2000: An open large scale video database for violence detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4183–4190. IEEE.

Ming Cheng, Kunjing Cai, and Ming Li. 2021b. Rwf-2000: an open large scale video database for violence detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4183–4190. IEEE.

Chunhui Ding, Shouke Fan, Ming Zhu, Weiguo Feng, and Baozhi Jia. 2014. Violence detection in video by using 3d convolutional neural networks. In *Advances in Visual Computing: 10th International Symposium, ISVC 2014, Las Vegas, NV, USA, December 8-10, 2014, Proceedings, Part II 10*, pages 551–558. Springer.

Srinath Doss et al. 2024. Comparative analysis of news articles summarization using llms. In *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, pages 1–6. IEEE.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Raisa Islam and Owana Marzia Moushi. 2024. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*.

Noussaiba Jaafar and Zied Lachiri. 2023. Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. *Expert Systems with Applications*, 211:118523.

Gurmeet Kaur and Sarbjeet Singh. 2024. Revisiting vision-based violence detection in videos: A critical analysis. *Neurocomputing*, 597:128113.

meta. 2024. [link].

Mohamed Mustafa. 2020. Real-Life Violence Situations Dataset. `https://www.kaggle.com/datasets/mohamedmustafa/real-life-violence-situations-dataset`.

Anuja Jana Naik and MT Gopalakrishna. 2017. Violence detection in surveillance video-a survey. *International Journal of Latest Research in Engineering and Technology (IJLRET)*, 1:1–17.

Mohamed Nejjar, Luca Zacharias, Fabian Stiehle, and Ingo Weber. 2025. Llms for science: Usage for code generation and data analysis. *Journal of Software: Evolution and Process*, 37(1):e2723.

ID Ohwosoro, AE Edje, and CO Ogeh. 2024. A hybrid assault detection system using random forest enabled xgboost-lightgbm technique. *Nigerian Journal of Science and Environment*, 22(2):177–189.

Bruno Peixoto, Bahram Lavi, Paolo Bestagini, Zanoni Dias, and Anderson Rocha. 2020. Multimodal violence detection in videos. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2957–2961. IEEE.

Francisco A Pujol, Higinio Mora, and Maria Luisa Pertegal. 2020. A soft computing approach to violence detection in social media for smart cities. *Soft Computing*, 24(15):11007–11017.

Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541.

Ambreen Sabha and Arvind Selwal. 2024. Towards machine vision-based video analysis in smart cities: a survey, framework, applications and open issues. *Multimedia Tools and Applications*, 83(22):62107–62158.

Hasan Ahmed Salman, Ali Kalakech, and Amani Steiti. 2024. Random forest algorithm overview. *Babylonian Journal of Machine Learning*, 2024:69–79.

Innovation & Technology GOV UK Department for Science. 2025. Online safety act: Explainer.

Sakib Shahriar, Brady D Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. 2024. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences*, 14(17):7782.

Luísa Shimabucoro, Sebastian Ruder, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. Llm see, llm do: Guiding data generation to target non-differentiable objectives. *arXiv preprint arXiv:2407.01490*.

M. Soliman, M. Kamal, M. Nashed, Y. Mostafa, B. Chawky, and D. Khattab. 2019. Violence recognition from videos using deep learning techniques. In *Proceedings of the 9th International Conference on Intelligent Computing and Information Systems (ICICIS'19)*, pages 79–84, Cairo, Egypt.

Grace Studer. 2017. Live streaming violence over social media: an ethical dilemma. *Charleston L. Rev.*, 11:621.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. 2024. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

Fath U Min Ullah, Mohammad S Obaidat, Khan Muhammad, Amin Ullah, Sung Wook Baik, Fabio Cuzzolin, Joel JPC Rodrigues, and Victor Hugo C de Albuquerque. 2022. An intelligent system for complex violence pattern analysis and detection. *International journal of intelligent systems*, 37(12):10400–10422.

Fath U Min Ullah, Amin Ullah, Khan Muhammad, Ijaz Ul Haq, and Sung Wook Baik. 2019. Violence detection using spatiotemporal features with 3d convolutional neural network. *Sensors*, 19(11):2472.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

Gokul Yenduri, M Ramalingam, G Chemmalar Selvi, Y Supriya, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, G Deepti Raj, Rutvij H Jhaveri, B Prabadevi, Weizheng Wang, et al. 2024. Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*.