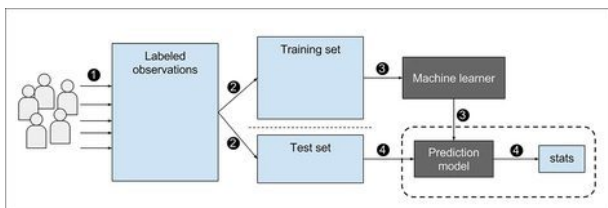


Predicting Violence in Advance With AI

Originally published Thursday 1st October 2026 at <https://www.unite.ai/predicting-violence-in-advance-with-ai/>



The template for a typical 'Can AI automate this?' line of inquiry is to feed multiple examples of something whose characteristics you want to isolate, and see if a model trained on that data can pick out that characteristic on similar data that was *not* used in training:



Schema for a 'Hail Mary' AI project: labeled examples are divided into training and test sets, with the model learning from the first and then being challenged with unseen examples from the second. If it has genuinely learned the characteristic of interest rather than memorized the training data, it should recognize the same pattern in the test set. [Source](#)

It's a bit of a 'Hail Mary', since such experiments rest on the assumption that the characteristic will eventually stand out in the data if the AI keeps looking at the dataset, and that there is actually any kind of 'underlying trend' to elicit. If the characteristic was something we could easily recognize, we either wouldn't need the AI model at all, or we would only be using it for routine automation.

This is not generally the case; most such research initiatives are 'fishing expeditions' for latent patterns that only superhuman repetition of attention could learn to recognize.

Often enough, the pattern turns out to be elusive, and may in fact be non-existent, because some events and traits may truly operate at pure chance; or else, the data is in some way inadequate, poorly curated, or poorly-trained.

Looking Ahead

In some cases, the prize is so desirable (cure cancer, nuclear fusion, overcome laws of thermodynamics, etc.) and so 'politically attractive' that the literature will return to it persistently, to see if better data or better techniques can yet elicit a characteristic pattern.

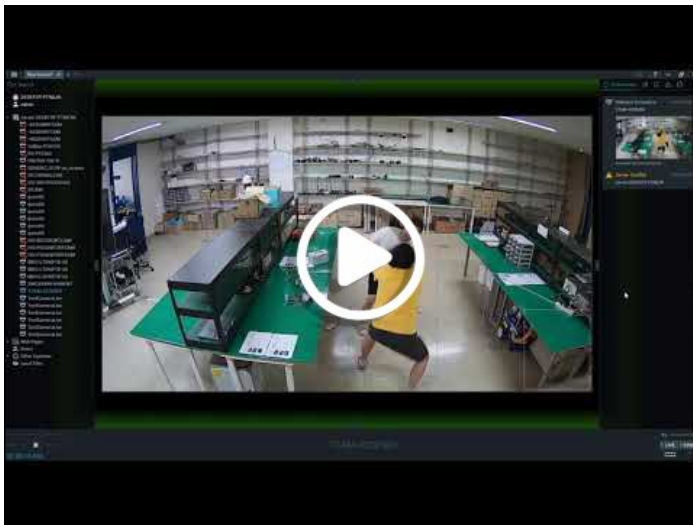
One such field is [Automated Video Understanding](#), as it relates to public spaces and police-involved domains. For instance, this year a new study [used real-world video footage of suicide attempts](#) to identify 'dwelling' trends prior to a person's suicide attempt, and to create a predictive system that could spot the characteristic way that they move around a public train platform prior to jumping:



From the 2026 paper 'Suicide Risk Assessment from AI-powered Video Surveillance: An Interpretable Framework for Prevention in Metro Stations', predictions from genuine rail suicide attempt and the other a non-incident case. Heatmaps beside each frame indicate areas of higher and lower risk on the platform, based on a pe tunnel mouth and patterns observed in previous suicide attempts. [Source](#)

In this case, the 'dwell patterns' (hard left and right in the image above) typical of a person conceiving of jumping in front of a train appeared to have a distinct signature that could be used as a pattern-recognition alert system in automated surveillance circuits.

Another popular strand of research in this line is *violence detection/recognition*, where unattended surveillance systems could be made capable of recognizing violent action, in order to bring the matter quickly to human attention – and even to trigger ancillary operations, such as flagging the footage [for long-term preservation](#), and ensuring that recording [is actually taking place](#), among other measures.



A (non-real) test for a violence detection system from IVS. [Source](#)

According to [new research](#) from the University of Alabama, the literature to date has concentrated on recognizing violence when it is already happening. For the new work, the researchers trained a model only on video footage *prior* to a violent event (or to no event, for statistical balance), to see if there might be recurrent and trainable 'pre-violent' characteristics – in this case, a combination of body motion, facial disposition, and audio, which together could provide a predictive signal for violent action, instead of just responding to it once it has occurred.



Examples of 'risky' pre-incident sequences from the study 'Can We Anticipate Violence? Multimodal Learning from Pre-Incident Behavioral Cues', shown across fc system tracks body joints and movement over time, looking for changes in posture, speed and interaction patterns that may distinguish escalating situations before [Source](#)

Of course, such work is [likely to draw concern](#), if developed or implemented. As I have [noted before](#), fictional outings such as the TV show *Person of Interest* and the 2002 sci-fi actioner *Minority Report* have, among many similar outings, reflected popular fears around predictive technologies in the hands of police or regulatory authorities.

As it stands, in terms of predicting violent action from body, facial and audio cues (rather than factors [such as context](#), time of day location or historical crime figures), it's a moot point, as this is a very fresh line of research, according to the new paper's authors.

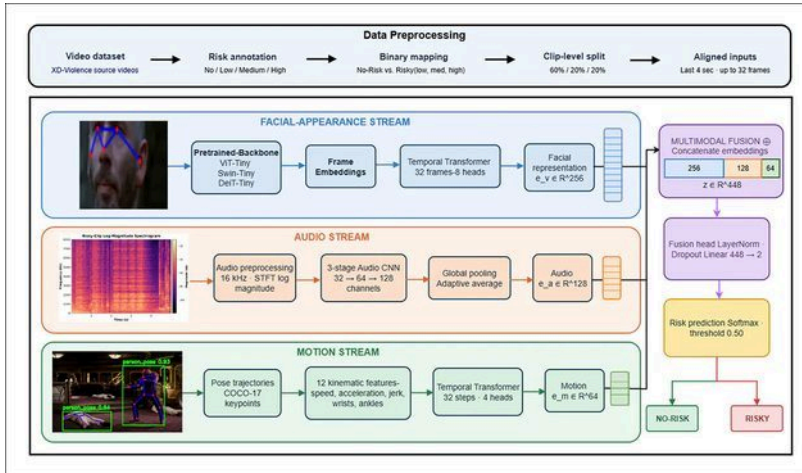
In tests, the researchers compared facial appearance, audio and body-motion cues separately and in combination, using three different vision backbones. The strongest results came when all three signals were combined, suggesting that the predictive effect was not being driven by any single type of cue:

'The best configuration, DeiT-Tiny with audio, facial appearance, and motion, achieves 91.21% accuracy, 88.96% balanced accuracy, 93.65% F1-score, and 96.38% ROC-AUC on the held-out test set.'

'These results suggest that complementary appearance, acoustic, and kinematic cues provide useful evidence for recognizing elevated pre-incident risk.'

Approach

The dataset curated for the study is derived from the [XD-Violence](#) collection, and processes three modalities in concert: facial appearance, audio, and pose-estimated motion:



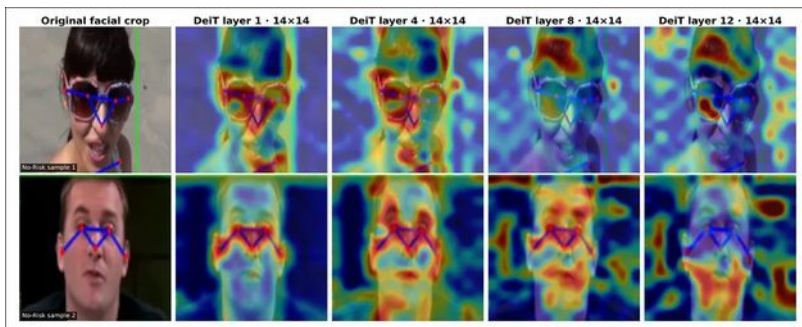
The study's processing pipeline. Each clip is labeled by risk level, reduced to 'no-risk', or 'risky', and then split into training, validation and test sets. Up to four seconds through three branches: facial appearance across frames, audio converted into a spectrogram, and body motion derived from tracked keypoints such as speed and facial, audio and motion features are combined into one representation and passed to a final classifier, which assigns the clip a probability of being no-risk or risky.

The resulting curated set contained 443 pre-incident clips (we'll discuss this number at the end of the article).

The videos selected from the XD-Violence collection, the paper reports, were chosen as suitable subjects for the kind of human-focused analysis the project was aiming to implement. Conversely many similar prior approaches had concentrated on the *entirety of the scene* as a predictive context.

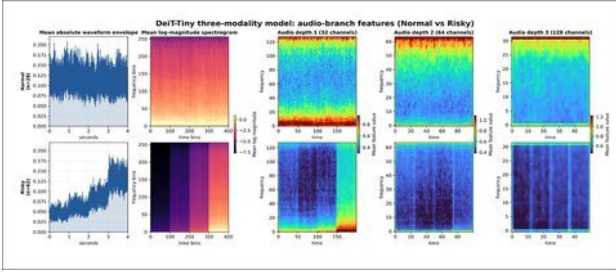
Clips used did not exceed four seconds, and were sampled at 8fps. A single-layer [Transformer](#) with eight [attention heads](#) was used to collate the resulting 32 frame samples, while learned [positional embeddings](#) maintained the frame order. The backbone remained [frozen](#) during multimodal training, in order to avoid [overfitting](#).

Sample [activation maps](#) (including facial images) are shown earlier in the article, for 'risky' activations, and below, for No-Risk samples:



No-risk examples from the study's facial-analysis branch, complementing the risky motion examples shown earlier. The original facial crop appears at left, followed progressively deeper layers of the DeiT-Tiny vision model. Warmer areas indicate stronger feature responses, showing how the model's attention shifts and becomes more focused on specific facial features as it progresses through the network.

The audio components derived from the source clip were resampled to 16kHz and conformed through normalization and alignment with the corresponding visual window, before being converted into standardized [log-magnitude spectrograms](#) using a 512-point short-time [Fourier transform](#), and a 400-sample [Hann window](#).



Comparison of average audio patterns for normal and risky pre-incident clips. The top row shows normal clips and the bottom row risky clips, progressing from the original waveform and spectrogram to features learned at three successive network depths. Risky clips show greater variation over time, including increased acoustic activity toward the end of the four-second window, while the network progressively converts these differences into higher-level features.

For the motion component, body movement was reduced to a set of numerical descriptors derived from [COCO-17](#) keypoints detected by a [YOLO](#) pose estimator.

Twelve measures were calculated, including *average* and *maximum speed*, *acceleration* and *jerk*, *variation in speed*, *wrist and ankle speeds*, and overall *movement energy*. These values were normalized for body size, so that larger people would not automatically register as moving more. Additionally, measurements from multiple people in the same clip were combined.

After [standardization](#) against the training data, the motion features were expanded into a 64-[dimensional representation](#), and passed through a temporal Transformer, allowing changes in movement across the clip to be considered as a sequence, rather than as isolated frames.

The three streams were then fused into a single representation, with facial appearance given the largest share, followed by audio and motion. After normalization and dropout, a final classifier converted that combined representation into probabilities for 'normal' and 'risky', using 0.5 as the cutoff.

For the facial appearance stream, three [ImageNet-pretrained](#) backbones were used: [Swin-Tiny](#); [ViT-Tiny](#); and [DeiT-Tiny](#).

Tests

For testing, clips were manually assigned one of four labels: *No Risk*, *Low Risk*, *Medium Risk* or *High Risk*, based on visible behavior, motion and audio. For the final binary task, the three risk categories were grouped together as 'risky', while *No Risk* became 'normal'. To reduce leakage between training and evaluation, clips taken from the same source video were kept within the same dataset partition.

The violent incident itself was removed from every risky clip, leaving only the preceding behavior for the model to judge:

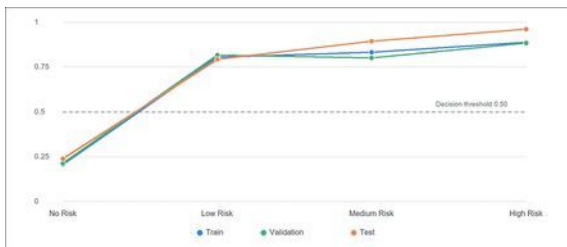
Ablation	Model	Accuracy	Bal. Acc.	Precision	Recall	F1	ROC-AUC
Motion only	Temporal Transformer	82.42	76.08	82.86	93.55	87.88	86.21
Audio only	Audio CNN	67.03	73.05	92.11	56.45	70.00	78.75
Facial appearance only	ViT-Tiny	87.91	85.62	90.48	91.94	91.20	91.94
Facial appearance only	Swin-Tiny	89.01	87.35	91.94	91.94	91.94	95.22
Facial appearance only	DeiT-Tiny	87.91	85.62	90.48	91.94	91.20	94.22
Audio + Motion	Audio CNN + Temporal Transformer	78.02	75.61	85.00	82.26	83.61	88.71
Facial appearance + Motion	Swin-Tiny + Temporal Transformer	84.62	83.20	90.00	87.10	88.52	93.83
Audio + Facial appearance	ViT-Tiny	86.81	83.90	89.06	91.94	90.48	93.10
Audio + Facial appearance	Swin-Tiny	89.01	86.43	90.62	93.55	92.06	96.11
Audio + Facial appearance	DeiT-Tiny	87.91	84.71	89.23	93.55	91.34	93.66
Audio + Facial appearance + Motion	ViT-Tiny	86.81	84.82	90.32	90.32	90.32	94.33
Audio + Facial appearance + Motion	Swin-Tiny	90.11	87.24	90.77	95.16	92.91	96.11
Audio + Facial appearance + Motion	DeiT-Tiny	91.21	88.96	92.19	95.16	93.65	96.38

Test-set results showing how performance changes when facial appearance, audio and motion are used separately or in combination across the three visual back isolating the contribution of each input stream. Higher scores mean better performance, and bold values denote the best result in each metric.

As already noted, the strongest overall result came from DeiT-Tiny with all three modalities combined. The ablations make the contribution of each stream clearer: facial appearance was already a strong predictor on its own, while motion performed moderately and audio was considerably weaker in isolation.

The best results were obtained when facial, audio and motion information were combined, indicating that the weaker streams still added useful complementary information to the visual signal.

Although the model was trained only to distinguish 'normal' from 'risky', its outputs were also compared against the original four-level labels: *No Risk*, *Low Risk*, *Medium Risk* and *High Risk*. As shown below, predicted risk rises broadly in step with those categories across the training, validation and test sets:



Predicted risk scores across the four original risk categories. Although the model was trained only to separate normal from risky clips, scores generally rise from *No Risk* through *High Risk* across the training, validation and test sets. The dashed line marks the 0.50 decision threshold.

The authors comment:

'This result is important because the model is never explicitly trained to separate Low, Medium, and High Risk. Even so, its predictions follow the progression of the original risk levels.'

'This suggests that the model is capturing meaningful changes in pre-incident behavior, rather than only learning a simple Normal-versus-Risky boundary. The result also provides evidence that useful behavioral cues can appear before the annotated incident begins.'

And they conclude:

'These findings suggest that useful discriminative information can be present in the period preceding an incident, highlighting the potential of moving violence analysis beyond event-present detection toward earlier risk recognition.'

Conclusion

Did the researchers really find a 'pre-violence' trait combination? Well, with only 443 clips used in the experiments, overfitting is a notable risk, and something multimodal models and pretrained vision backbones are especially prone to. The study partly mitigates this by freezing the visual backbone, separating source videos across train/validation/test splits, and testing multiple modality combinations.

Strong performance on a small dataset doesn't prove the underlying pattern is obvious or even extant, since it can also reflect dataset-specific cues, annotation bias, source leakage not captured by the split, or an unusually easy test set.

But as it stands, the result is intriguing precisely *because* the dataset is so small, and because overfitting seems to have been adequately protected against. However, the findings would need replication on a larger dataset before the learned pattern could be considered 'robust'.

First published Thursday, October 1, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Website: martinanderson.ai

Contact: martin@martinanderson.ai