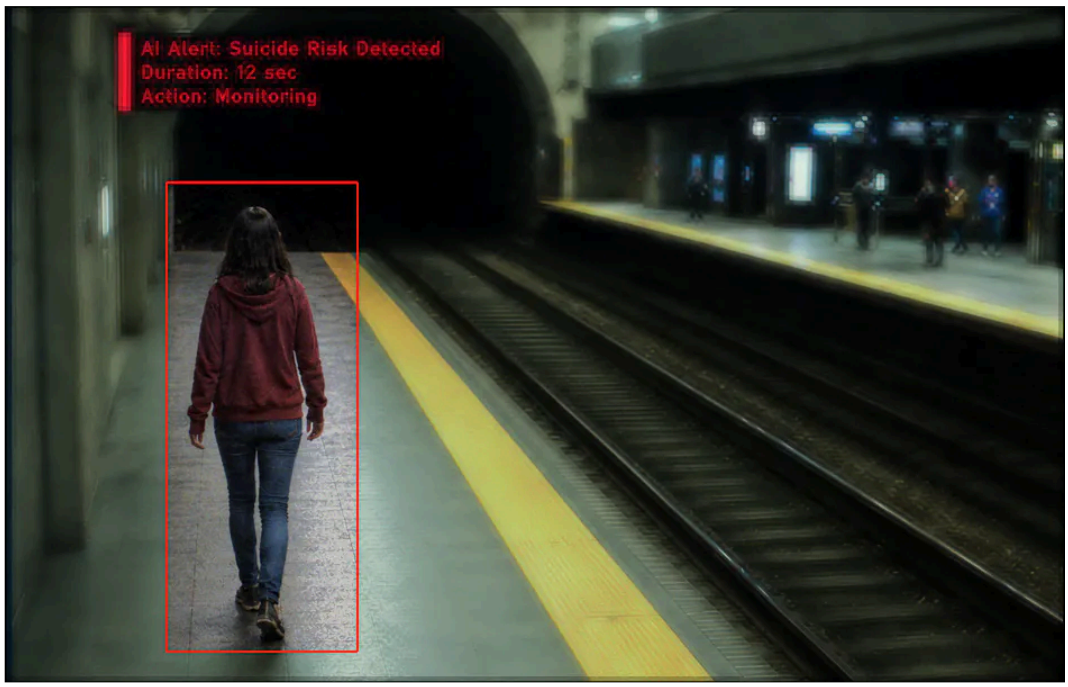
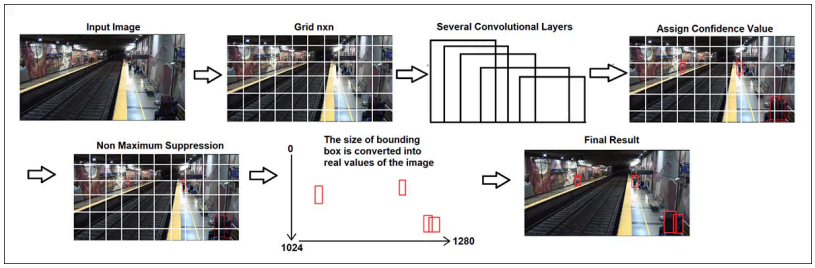
Anticipating and Preventing Metro Platform Tragedies With AI

Originally published at https://www.unite.ai/anticipating-and-preventing-metro-platform-tragedies-with-ai/



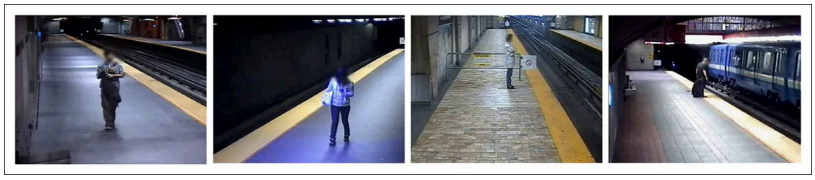
An AI system trained on real subway surveillance footage claims it can spot the warning signs of a suicide attempt minutes before it happens, tracking behaviors such as pacing, lingering at the platform edge, and repeatedly looking into the tunnel.

Machine learning systems have been trialed as platform event-monitoring systems for some years, usually with some variation of the popular You Only Look Once (YOLO) series of image recognition applications powering scenarios where pedestrians may have fallen, or a crime is being committed, or where the station platform is simply overcrowded (allowing station authorities to regulate access and remediate the problem).



From the 2024 paper 'Train Station Pedestrian Monitoring Pilot Study Using an Artificial Intelligence Approach', the stages by which YOLOV7 identifies passengers

With an increase in the number of attempted or successful rail suicides over the past 3-5 years (in regions such as the UK, Canada and the Netherlands), interest has increased in the potential of machine learning systems to identify suicide-inclined behavior on rail and metro platforms, based on disposition and diverse other factors:



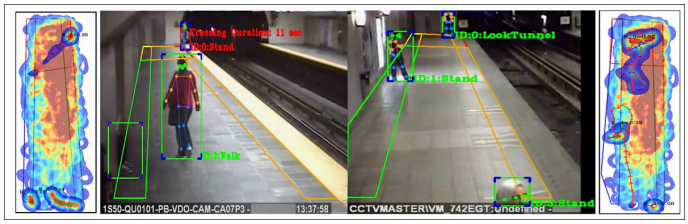
Dwellers on the threshold: example data from the dataset powering the STARR project, which features in the new paper under discussion in this article. Source

In aggregate, the variety of projects looking to leverage AI for suicidal platform-based behavior have not, to date, adopted a uniform methodology or underlying system or common approach – not least because the methods that power such systems are constantly evolving, along with the psychological and psychiatric knowledge that brings insight to this kind of AI surveillance.

Cutting Edge

Now, a new study from Canada offers a proposed formalization of this strand in the research literature, as *Suicide Risk Assessment* (SRA), in the context of suicide attempts at metro stations.

In collaboration with the Montreal transport authorities, the researchers involved in the new study gained access to footage of 66 real-world suicide attempts, as captured by platform cameras within the authorities' purview:



From the new paper, output predictions from two frames one depicting a genuine rail suicide attempt, and the other not. On either side of each image is depicted a heatmap of dangerous and safer areas in the platform under surveillance, depicting in each case a person's 'dwell tendency' in regard to the tunnel mouth, interpreted through historic knowledge of the tendencies of real 'jumpers'. Source

Though it was necessary to artificially address the class imbalance that occurs with such a limited ground truth dataset, this is nonetheless rare data at some viable volume; one could hope that future projects from transit authorities around the world might allow for a multi-country dataset with a higher volume of examples. However, understandably, the extremely sensitive nature of such footage makes this more than a casual or easy prospect.

The initiative, the authors contend, is the first to coalesce the diverse tasks that define the pursuit into a schema, and brings with it a new benchmark for the metro platform suicide scenario.

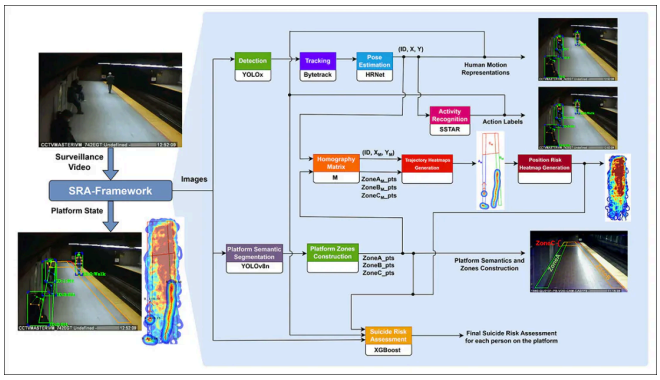
The authors state:

> *'Unlike approaches that focus on isolated subtasks or attempt to infer intent directly, our formulation assesses suicide risk from accumulated evidence by incorporating person tracking, activity recognition, semantic segmentation of the platform, and trajectory-driven risk heatmap modeling.*
>
> *'By formalizing SRA as a distinct task and benchmarking a complete operational pipeline achieving 83.2% ROC-AUC on real surveillance data, this work highlights the complexity of suicide risk assessment and opens new directions for research on interpretable AI systems for social good.'*

The new paper is titled *Suicide Risk Assessment from AI-powered Video Surveillance: An Interpretable Framework for Prevention in Metro Stations*, and comes from four authors across Université TÉLUQ, Polytechnique Montréal, and the Université du Québec à Montréal.

Method

The authors' framework analyzes a live surveillance video feed to generate a continuously-updated suicide-risk score for each tracked passenger. Individuals are detected, tracked, and converted into simplified body-pose representations, after which a skeleton-based activity-recognition system identifies actions over short periods of time:



System pipeline for estimating passenger suicide risk from surveillance video, showing how tracking, pose estimation, activity recognition, platform zoning, and trajectory heatmaps are combined to convert individual movements and behaviors into a continuously updated risk score for each person on the platform.

The platform is then divided into meaningful zones, allowing movement patterns – such as repeated pacing between different areas – to be detected. Passenger trajectories are projected onto a map of the platform, making it possible to generate heatmaps that highlight the areas most frequently-occupied, or crossed by people associated with elevated risk.

Finally, the system cross-references these spatial patterns against observed behaviors to produce an individual suicide-risk assessment for each person on the platform – a process the authors dub *risk inference*.

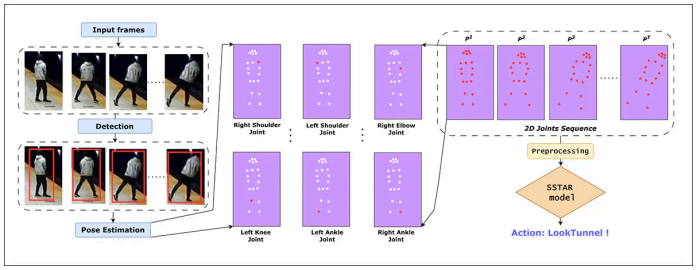
The authors used a pretrained YOLOX implementation as the human detector for their system, finding that its out-of-the-box state is perfectly usable for this purpose. ByteTrack was used to orchestrate multi-object tracking.

Each person individuated by these processes is assigned their own pretrained HRNet model, providing joint estimation and 2D body keypoints inside a bounding box determined by the outermost of these estimations:

Examples of joint estimation from HRNet, used in the new project. Source

The poses evaluated from video data from the metro platform are built up into cumulative maps defining historical movement (see the 'platform heat maps' at the side of the earlier image above).

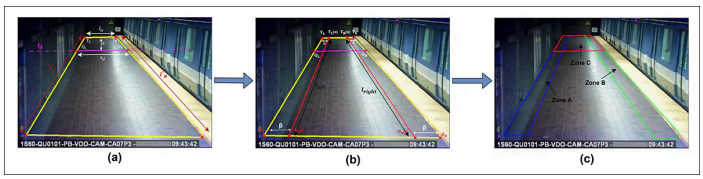
The new system incorporates the STARR framework, a prior work designed to evaluate the probability of suicidal behavior at platforms:



Pose estimation from the STARR framework. Source

In this case, STARR is used to detect three self-explanatory passenger action annotations: *LookTunnel*; *Walk*; and *Stand*.

To incorporate environmental context, the system's conception of the platform is divided into semantically meaningful zones using a YOLOv8n semantic-segmentation model trained on manually-annotated platform images:



Platform semantics: the zoning process used by the system to convert a segmented platform into three behaviorally meaningful regions. The resulting wall-proxim yellow-line-proximal, and tunnel-adjacent zones provide the spatial context used to evaluate passenger movements and risk-related behaviors.

The resulting segmentation map is used to estimate the platform boundaries and define three operational areas: a Wall-Proximal Zone near the platform wall; a Yellow-Line Proximal Zone where passengers can approach the platform edge while remaining within safety boundaries; and a Platform Far-End Zone nearest the tunnel entrance.

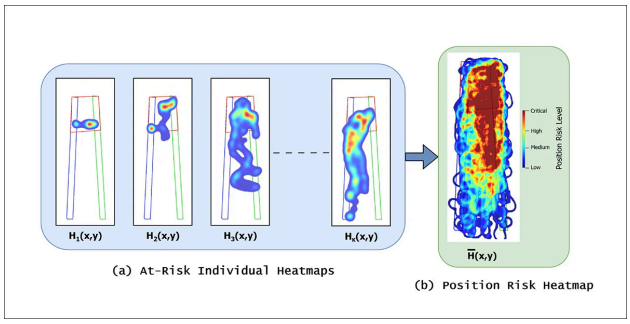
These zones provide the spatial context needed to identify behaviors that psychological studies have associated with elevated suicide risk. In particular, they allow the system to detect *repeated movement between the wall and the yellow line*, along with entry into the far-end area of the platform.

Combined with the trajectory heatmaps generated earlier, these spatial indicators are later incorporated into the final suicide-risk assessment.

Interestingly, the paper notes that one hallmark of suicide attempts is the tendency to leave an object on the platform; however, the authors were not able to incorporate this into this version of the project, leaving it for future work.

A Map of Platform Risk

Rather than relying only on the behavior of a single person, the framework also combines trajectory heatmaps from multiple known at-risk cases to create a platform-wide 'position risk heatmap':



Building a platform risk map from the movements of multiple at-risk passengers. Areas that repeatedly attract lingering, pacing, or other risk-associated behavior become increasingly prominent and are later used as a factor in the final risk assessment.

Areas that repeatedly attract prolonged occupancy emerge as higher-risk regions, while locations associated with brief or infrequent visits remain lower-risk. The resulting position-risk score becomes one of the inputs used in the final suicide-risk assessment.

The final risk score is based on eight indicators accumulated over time: a position-risk score derived from the platform heatmaps; whether a passenger walks or stands on the yellow line; the number of yellow-line crossings; the total time spent on the yellow line; the longest uninterrupted period spent on the yellow line; repeated back-and-forth movement between the wall and the yellow line; repeated orientation toward the tunnel; and entry into the tunnel-adjacent end of the platform.

These behavioral and spatial signals are then combined via an XGBoost model, producing a continuously updated suicide-risk estimate for each individual on the platform.

Data and Tests

Tests were conducted on surveillance footage supplied by the Société de transport de Montréal (STM), comprising 66 five-minute recordings captured before real suicide attempts, together with 56 matched control recordings from the same cameras, at comparable times when no suicide attempt occurred.

With the assistance of psychology and experts in suicidal behavior, individual passengers were annotated according to whether they appeared in an at-risk or control scenario, producing a dataset of 256 individuals, of whom 66 were associated with suicide-attempt cases, and 190 assigned to the control group.

To prevent information leakage, all individuals extracted from the same recording were assigned to the same data split, with 75% of the data used for training and 25% reserved for testing while preserving the balance between at-risk and control cases.

The XGBoost classifier was trained for 300 boosting iterations, at a learning rate of 0.05, with subsampling for both training instances and features, to improve generalization. Because the dataset contains substantially more control cases than at-risk cases, the training process compensated for this by assigning additional weight to the minority class.

Performance was evaluated primarily using the Area Under the Receiver Operating Characteristic Curve (ROC-AUC), measuring how effectively the system distinguishes between at-risk and control individuals.

Additional metrics comprised *sensitivity*, measuring correct identification of at-risk cases; *specificity*, measuring correct identification of control cases; *false-positive rate*, reflecting false alarms (FPR); and *false-negative rate*, reflecting missed detections (FNR). A deliberately low decision threshold was adopted to prioritize early identification of potentially at-risk situations:

Configuration	Modules		Statistical Measures (± SD)				
	Detection / Tracking	*Act. Rec.*	*ROC-AUC*	*Sensitivity*	*Specificity*	*False Alarm Rate (FPR)*	*Missed Detection Rate (FNR)*
GT-Assisted	YOLOx / Bytetrack	GT	0.893 ± 0.051	0.631 ± 0.228	0.913 ± 0.056	0.087 ± 0.056	0.369 ± 0.228
GT-Assisted	GT	SSTAR	0.919 ± 0.054	0.800 ± 0.187	0.915 ± 0.046	0.085 ± 0.046	0.200 ± 0.187
Upper Bound	GT	GT	0.924 ± 0.055	0.867 ± 0.113	0.929 ± 0.064	0.071 ± 0.064	0.133 ± 0.113
Fully Automatic	YOLOx / Bytetrack	SSTAR	0.832 ± 0.095	0.631 ± 0.228	0.902 ± 0.071	0.098 ± 0.071	0.369 ± 0.228

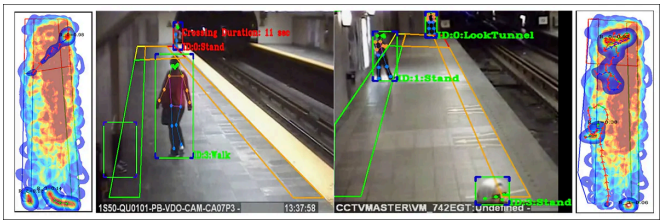
Performance of the suicide-risk assessment framework under four configurations, comparing ground-truth-assisted, upper-bound, and fully automated variants acr ROC-AUC, sensitivity, specificity, false-alarm rate, and missed-detection rate – and showing the performance impact of replacing ground-truth annotations with automated detection, tracking, and activity-recognition modules.

As seen in the initial results table above, the fully automated system achieved a ROC-AUC of 0.832, while replacing the automated detection and tracking components with ground-truth annotations increased performance to 0.919.

Replacing only the activity-recognition module produced a smaller gain, raising ROC-AUC to 0.893. The highest score, 0.924, was obtained when all inputs were supplied from ground-truth annotations, suggesting that detection, tracking, and downstream indicator extraction remain the main sources of error in the current pipeline.

Analysis of the trained XGBoost model indicated that *direct interaction with the yellow line* was among the strongest predictors of elevated risk, followed by the *number* of yellow-line crossings, and repeated back-and-forth movement between platform zones. Time spent on the yellow line, and the position-risk score, also contributed substantially, while tunnel-directed attention and entry into the platform's far-end zone provided additional, but less decisive signals.

Returning to the qualitative results previewed above, the framework assigned high risk scores to individuals later associated with suicide attempts while assigning substantially lower scores to surrounding control passengers:



As seen earlier in the article, the qualitative results shown in the paper feature the output predictions of two images from surveillance data, with heatmaps either side to indicate dwelling and other platform activities.

In one case, a risk score of 0.98 was linked to prolonged presence on the yellow line and occupancy of regions identified as high-risk by the position-risk heatmap. In another, an at-risk individual received a score of 0.92, while nearby control passengers received much lower estimates.

According to the authors, these distinctions emerge from the accumulation of multiple indicators, rather than any single behavior. Prolonged crossing of the yellow line, repeated orientation toward the tunnel, and sustained presence in high-risk areas of the platform all contribute to elevated risk estimates.

The authors conclude:

> *'Beyond performance, our study emphasizes interpretability, showing that risk assessments are driven by intuitive indicators aligned with established behavioral and spatial risk factors.*
>
> *'This positions the proposed framework as a meaningful bridge between AI-based surveillance systems and interdisciplinary research on suicide prevention.'*

Conclusion

On a personal note, it is an increasingly rare relief to find an AI paper worth reporting that is not likely to create an incendiary reaction in some part of the populace, since it would be hard to dispute the value of the objectives behind this kind of project.

On a practical note, the very small amount of pixels occupied by the head, and the relatively small amount of screen space occupied by the entire person under surveillance in this scenario, make it very difficult to tell if the individual is looking frequently at the tunnel – one of the telltale signs of the potential rail suicide.

As ever, in projects regarding surveillance infrastructure, this would appear to be an issue around resolution and resources: if there were more cameras at more frequent intervals covering the platform, including one specifically covering the tunnel exit (i.e., the tunnel aspect from which a metro train suddenly appears), there would be scope to involve some of the various and constantly developing frameworks around gaze direction. As it stands, the current work relies on evaluating the entire direction of the body to signal that the subject is regarding the tunnel.

In the end, the issue is a budgetary one, at least as far as rail infrastructure is concerned; if all platforms were outfitted with barriers and gates – features which show up infrequently in London Underground stops, and in the metro networks of other cities around the world – then the platforms would offer no opportunity for self-harm.

For sure, increased surveillance is the cheaper option, and early identification of characteristic signs of self-harm could allow direct intervention before tragedy occurs.

First published Tuesday, June 9th, 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.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai