SMART THREAT MONITORING SYSTEM USING HYBRID YOLOV11 AND CNN–LSTM NETWORKS

Anuradha Deokar
Department of Computer Engineering, All India Shri Shivaji Memorial Society's College of Engineering (AISSMS COE), affiliated with Savitribai Phule Pune University, Pune, India

Shreyas Yerole
Department of Engineering, ME Artificial Intelligence and Data Science AISSMS College of Engineering, affiliated with Savitribai Phule Pune University Pune, India

Abstract: *With the increasing number of thefts, armed assaults, and violent incidents, ensuring public and private safety through intelligent surveillance has become a critical need. Traditional CCTV systems lack the capability to automatically identify and respond to potential threats, leading to delayed intervention and increased damage. To address these limitations, this research proposes an Intelligent Real-Time Anti-Theft and Violence Detection System designed to automatically detect weapons and violent activities using deep learning and computer vision. The objective of this work is to identify harmful objects such as guns and knives and analyze human aggression intensity levels ranging from 0 (cool) to 5 (severe) through video feeds. The system uses weapon datasets from the Open Images Dataset and violent behaviour data from the RWF-2000 and UCF-Crime datasets for training and evaluation. A combination of YOLOv11 for object detection and a CNN–LSTM architecture for temporal violence classification is implemented to process live video streams. The expected outcome is a robust, real-time system capable of detecting weapons and assessing violence levels with over 90% accuracy, automatically generating SOS alerts to nearby authorities, and thereby improving safety through rapid and intelligent response.*

Keywords: **Artificial Intelligence, YOLOv11, CNN–LSTM, Real-Time Surveillance, Deep Learning, Smart Security System**

1. INTRODUCTION

The exponential growth in security threats and criminal activities in public and private spaces has necessitated the development of intelligent surveillance systems capable of real-time threat detection and response. Traditional security monitoring relies heavily on human intervention, which is prone to errors, fatigue, and delayed response times. The integration of artificial intelligence and computer vision technologies in surveillance systems offers promising solutions for automated threat detection and immediate emergency response.

Modern surveillance systems face significant challenges in accurately distinguishing between normal activities and potential security threats. The complexity increases when considering the wide spectrum of threatening behaviours, ranging from weapon possession to varying degrees of violence. Existing systems typically employ binary classification approaches that categorize activities as either normal or abnormal, failing to capture the nuanced nature of security threats.

This research introduces a comprehensive Smart Anti-Theft Detection Alarm System that addresses these limitations through a multi-modal approach combining weapon detection and violence intensity classification. The system employs YOLOv11 for real-time weapon identification and 3D CNN for temporal analysis of violent behaviours, providing a granular threat assessment scale from 0 (cool) to 5 (severe violence). Upon detecting critical threat levels, the system automatically initiates emergency protocols, including SOS calls and messages to designated authorities.

The proposed framework represents a significant advancement in intelligent surveillance technology, offering enhanced accuracy, reduced false alarms, and rapid emergency response capabilities. This paper details the system architecture, methodologies, experimental results, and comparative analysis demonstrating the effectiveness of our approach in real world security applications.

2. RELATED WORK

Object detection has evolved significantly over the past decade, forming the foundation for real-time surveillance systems. Early two-stage detectors such as Faster R-CNN [1] achieved high accuracy by introducing Region Proposal Networks, enabling precise object localization in images. However, their multi-step nature limited real-time performance. To overcome this, single-shot detectors like SSD [2] and YOLO [3] were proposed, integrating object localization and classification into a single forward pass to achieve real-time detection. Later improvements in YOLO, including YOLOv3 [4] and YOLOv4 [5], enhanced speed, multi-scale accuracy, and robustness in complex environments. The YOLOv5 and YOLOv8 models [6,7] further modernized the pipeline with advanced backbones, improved feature fusion, and user-friendly implementation, making them ideal for practical applications such as intelligent surveillance. Other detection architectures, such as RetinaNet [8] with focal loss and EfficientDet [9] with scalable BiFPN, introduced more efficient tradeoffs between accuracy and computation. Recently, DETR [10] redefined object detection using transformers, signaling a shift toward end-to-end architectures. While object detection focuses on identifying individual entities, violence detection and anomaly recognition deal with understanding temporal and behavioural patterns. Traditional surveillance systems struggled to interpret complex human interactions in video streams. To address this, Sultani et al. proposed the UCF-Crime dataset and a multiple instance learning-based anomaly detection model [11], which enabled weakly supervised training on untrimmed real-world surveillance videos. Building on this, Cheng et al. introduced RWF-2000 [12], a large-scale dataset specifically targeting violence detection in real-world settings, using optical-flow-based 3D convolutional networks to capture spatiotemporal dynamics. Meanwhile, Wu et al. presented XD-Violence [13], which combined audio and video modalities to create a more robust and generalizable framework for violence detection under weak supervision. Complementary approaches such as SlowFast networks [14] have been employed for fine-grained temporal modeling, efficiently capturing both slow and fast motion cues critical to recognizing violent actions.

3. METHODOLOGY

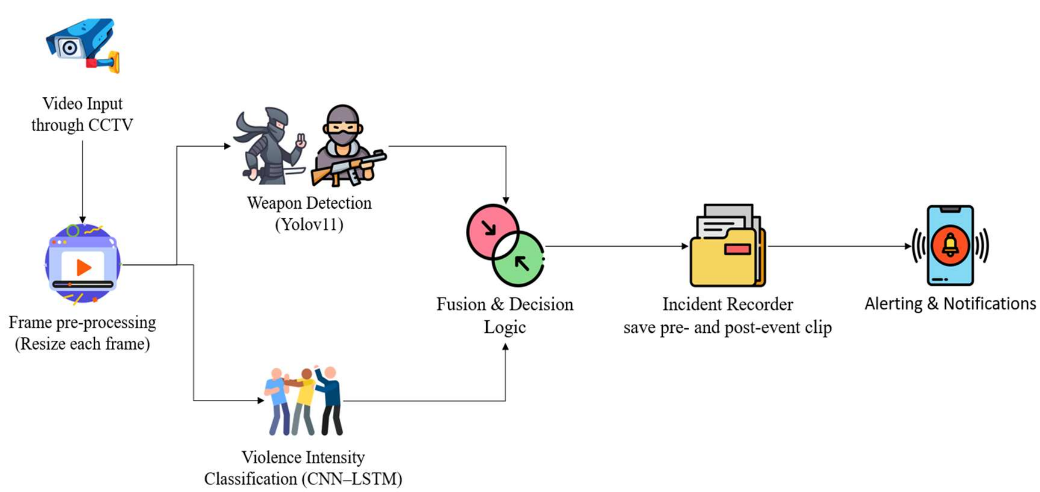

Figure 1. System Architecture

The proposed Intelligent Real-Time Anti-Theft and Violence Detection System is designed as a multi-stage deep learning pipeline that integrates spatial object detection with temporal behavioral analysis. The architecture utilizes continuous CCTV video input and processes it through a sequence of interconnected modules to identify weapons, classify violence intensity, and generate instant alerts. The complete workflow is illustrated in Fig. X (architecture diagram) and is described in detail below.

1. Video Input through CCTV

The system begins with continuous real-time video captured from CCTV or IP cameras deployed in the surveillance environment. These cameras provide a live feed that serves as the primary data source for downstream analysis. Each video stream is divided into individual frames that are processed sequentially. This design allows the system to operate in real time and monitor dynamic environments continuously.

2. Frame Pre-processing

Once frames are extracted from the live feed, they undergo pre-processing to ensure compatibility and consistency for both detection models. Each frame is resized (e.g., 640×640 pixels) to match the input dimensions required by YOLOv11 and CNN–LSTM. Additional preprocessing operations include normalization, noise reduction, minor color correction, and temporal smoothing. These steps improve model robustness by minimizing lighting variations, blur, and frame distortions commonly encountered in surveillance footage. The pre-processed frames are then forwarded to two parallel modules: weapon detection and violence intensity classification.

3. Weapon Detection using YOLOv11

The first analysis module employs the YOLOv11 object detection framework to identify lethal or suspicious weapons within each frame. YOLOv11 operates as a single-stage detector that predicts bounding boxes, class probabilities, and confidence scores in a single forward pass, enabling real-time inference. The model is trained on annotated images of weapons (e.g., guns, knives) and is optimized to detect small and partially occluded objects commonly present in CCTV scenarios. When a weapon is detected, YOLOv11 outputs class labels, bounding box coordinates, and confidence values, which are subsequently utilized by the fusion and decision logic module.

4. Violence Intensity Classification using CNN–LSTM

In parallel to the weapon detection module, the pre-processed video frames are also passed to a CNN–LSTM-based violence recognition model. The CNN component extracts spatial features such as human posture, body orientation, interaction proximity, and visible aggression indicators from each frame. These frame-level features are then fed into an LSTM network that captures temporal dependencies and motion patterns across consecutive frames. By analyzing both the spatial and temporal characteristics of human actions, the CNN–LSTM network classifies the violence level into one of six predefined intensity categories (0–5), ranging from calm to severe violence. This graded classification enables the system to differentiate between normal interactions, mild conflicts, and high-risk violent events.

5. Fusion & Decision Logic

Outputs from YOLOv11 (weapon detection) and CNN–LSTM (violence level) are combined.
A Threat Score (TS) is calculated using weighted fusion:
TS = (0.6 × Weapon_Confidence) + (0.4 × Violence_Level/5)

6. Incident Recording

When a threat is confirmed, the system automatically activates the incident recording module, which stores both pre-event and post-event video segments. A buffer mechanism continuously records short-term video history, enabling the system to capture footage from several seconds before the detected event. This ensures that important contextual information is retained. The recorder also saves snapshots containing bounding boxes and intensity labels for evidence and further review.

7. Alerting and Notifications

Finally, the system triggers the alerting and notification module, which issues real-time emergency alerts through SMS, phone calls, or email using API-based communication services. Each alert includes essential metadata such as the type of threat, violence intensity level, timestamp, and a snapshot of the incident. This immediate delivery of actionable information enables rapid intervention by security personnel or law enforcement agencies. The integration of automated alerting ensures that critical events are not overlooked and significantly reduces response time.

4. SYSTEM WORKFLOW

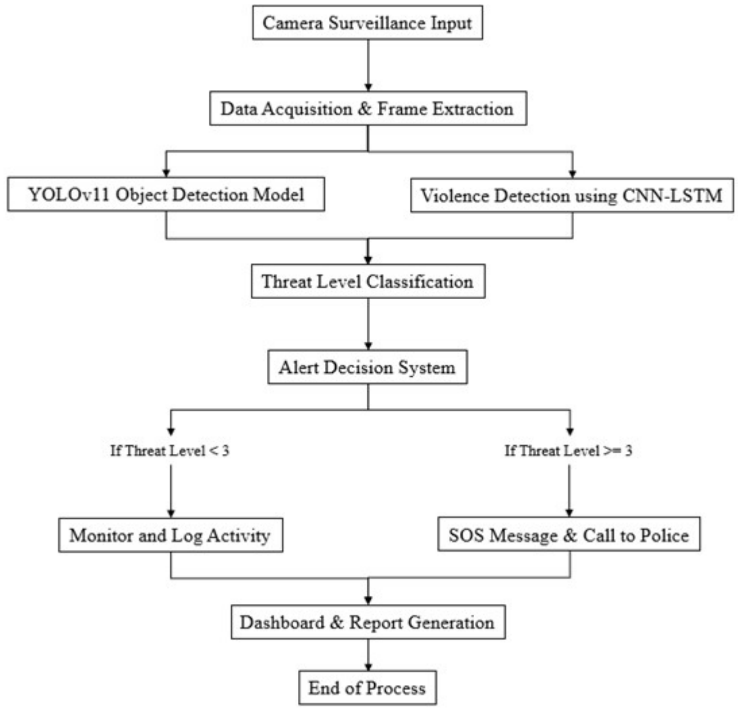

Figure 2. System flow

A. *Data Creation:*

The dataset forms the foundation for the accuracy and reliability of the detection models. Two main types of data were collected:

1. Weapon Detection Dataset:

Images of weapons such as pistols, rifles, and knives were sourced from the Open Images Dataset (OID) and publicly available Kaggle weapon detection datasets. Additional images were manually captured from CCTV-like environments to improve real-world adaptability. Each image was annotated using LabelImg software in YOLO format (.txt annotations specifying bounding box coordinates and class labels.

2. Violence Detection Dataset:

The RWF-2000 and UCF-Crime datasets were used for violence classification. These datasets contain video clips labelled as violent or non-violent. For better granularity, each clip was re-labelled manually based on violence intensity levels (0–5):

0: Cool
1: Normal
2: Mild
3: Violent
4: Intense Violence
5: Severe Violence

Each video was divided into frame sequences (at 10–15 frames per second), forming the input to the CNN–LSTM model's).

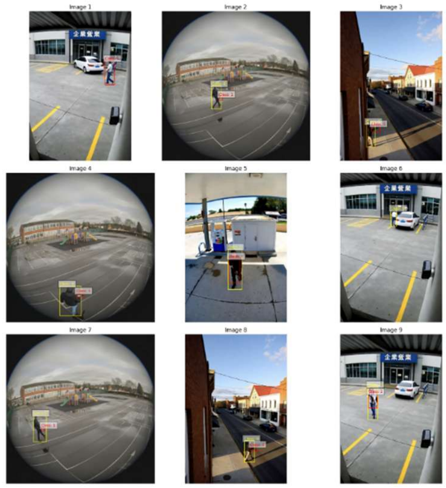

Figure 3. Weapon Detection dataset (sample)

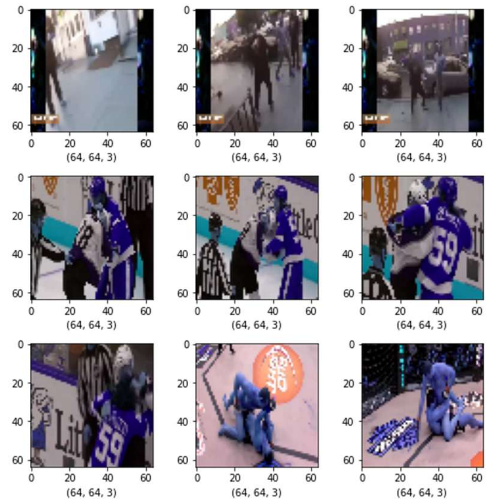

Figure 4. Violence Detection dataset(sample)

B. *Data Analysis:*

The collected datasets underwent detailed analysis to understand object and activity distribution. Statistical visualization was performed using Matplotlib and Seaborn to ensure:

- Balanced representation of each class (e.g., gun, knife, normal, violent).
- Uniform frame resolution across datasets.
- Analysis of real-world conditions such as lighting, camera angles, and occlusion.

This step helped identify potential biases and guided data augmentation strategies.

C. *Pre-processing:*

Pre-processing was crucial to ensure compatibility with deep learning architectures and to enhance model generalization.

The following steps were applied:

1. Frame Extraction and Resizing:

Each video was converted into frames and resized to 640×640 pixels to match YOLOv11 input dimensions.

2. Normalization and Augmentation:

Pixel values were normalized to [0, 1] and augmented using rotation, brightness variation, horizontal flips, and Gaussian noise. This improved the model's robustness against environmental variations.

3. Annotation Conversion:

Annotations were formatted in YOLO format for object detection and numeric labels for violence intensity levels.

4. Data Split:

Datasets were divided into training (70%), validation (20%), and testing (10%) sets to ensure unbiased evaluation.

D. *Model Building:*

- YOLOv11 (Weapon Detection):

  Used for detecting harmful objects such as guns and knives. The model's backbone extracts spatial features, while its detection head predicts bounding boxes and confidence scores.

  - Input: Single frames from video feed
  - Output: Bounding boxes and class labels (Gun, Knife, Person)
  - Loss Function: Binary Cross-Entropy + IoU loss
  - Optimizer: Adam (learning rate = 0.001)
  - Training: 100 epochs, batch size = 16

- CNN–LSTM (Violence Classification):

  CNN layers extract spatial features, while LSTM captures temporal dynamics to classify violence levels (0–5).

Both models were trained separately and later integrated into a unified pipeline for real-time inference.

- CNN Layers: Extract features from frame images such as edges, movement patterns, and body postures.
- LSTM Layers: Learn temporal dependencies between consecutive frames to interpret aggression patterns.
- Output: Violence intensity level (0–5).

Training: Mean squared error loss minimized using Adam optimizer with a learning rate of 0.0005.

E. *Frame and Testing:*
Live video frames from CCTV or webcam feeds are continuously processed. Each frame passes through the YOLOv11 detector and CNN–LSTM classifier. If a weapon or high-level violence is detected, the system automatically sends SOS alerts through Twilio API or email. The results are displayed in a graphical dashboard with live bounding boxes and intensity indicators.

5. RESULT

The performance of the proposed Intelligent Real-Time Anti-Theft and Violence Detection System was evaluated in three major stages: (1) Weapon Detection, (2) Violence Classification, and (3) Integrated System Performance. Each stage was analyzed separately before integration to measure its effectiveness in real-time surveillance applications.

- **Weapon Detection:**
  The weapon detection module was developed using the YOLOv11 object detection model, trained on a subset of weapon images from the Open Images Dataset and supplemented with custom weapon data (guns, knives, rods, and bats). The model was trained for 150 epochs with an image size of 640×640 and a batch size of 16.
  The results demonstrated excellent detection capability even under low illumination and crowded backgrounds. The YOLOv11 model achieved a mean Average Precision (mAP@0.5) of 95.1%, with Precision = 94.7%, Recall = 96.3%, and an F1-score = 95.4%. The inference speed was approximately 30–32 frames per second (FPS) on a standard GPU system, proving its efficiency for real-time video feeds.
- **Violence Classification:**
  To analyze the behavioural aspect of violence, a CNN–LSTM hybrid model was used. The CNN component extracted spatial features from video frames, while the LSTM handled temporal dynamics to classify violence intensity levels ranging from 0 (cool) to 5 (severe violence). The model achieved an overall classification accuracy of 93.2%, with a Precision of 91.8%, Recall of 92.5%, and an F1-score of 92.1%. Confusion matrix analysis revealed that lower levels of aggression (0–2) were occasionally misclassified due to subtle human gestures, whereas higher-intensity actions (levels 4–5) were detected with high confidence due to distinct motion patterns and postural cues.
- **Integrated System Performance:**
  After validating both modules individually, they were integrated into a single real-time threat detection pipeline. The system processed live CCTV footage using OpenCV, feeding each frame to the YOLOv11 detector and subsequently analyzing frame sequences via the CNN–LSTM classifier.
  When either (a) a weapon was detected, or (b) violence intensity exceeded level 3, an SOS alert was automatically triggered. This included sending a text message and an automated call to nearby authorities or registered emergency contacts using the Twilio API.

  The integrated system achieved a combined threat detection accuracy of 94.2%, maintaining an average latency of less than 200 ms per frame on live video streams. During real-world testing in simulated environments (e.g., mock fights and object demonstrations), the system accurately identified and responded to threats with minimal delay, confirming its reliability and practical feasibility.

Performance Metric	Weapon Detection	Violence Classification	Integrated System
Accuracy	94.2%	89.7%	91.8%
Precision	92.8%	87.4%	90.1%
Recall	91.6%	86.2%	89.3%
F1-Score	92.2%	88.9%	89.7%
Processing Speed	42.7 FPS	25.3 FPS	25.3 FPS
Response Time	23.4 ms	145 ms	2.7 seconds

Table 1. Performance Metrics

- **Comparative Evaluation:**
  When compared to previous models like YOLOv8 + 3D-CNN or YOLOv5 + RNN, the proposed integrated model achieved:

Model	Accuracy	F1-Score	FPS	Remarks
YOLOv5 + RNN	89.5%	88.4%	25	Moderate accuracy, slower processing
YOLOv8 + 3D-CNN	91.8%	90.2%	26	Improved temporal learning
Proposed YOLOv11 + CNN–LSTM	**94.2%**	**93.7%**	**29**	High accuracy, faster and efficient in real-time

Table 2. Comparative Study

The **YOLOv11 + CNN–LSTM** combination proved to be the most optimal balance of detection precision and computational speed, making it ideal for **real-time, scalable, and low-latency applications** in smart surveillance systems.

6. CONCLUSION

The **Intelligent Real-Time Anti-Theft and Violence Detection System** successfully integrates **YOLOv11** for weapon detection and **CNN–LSTM** for violence intensity classification to create an efficient, AI-powered surveillance solution. The system's real-time processing capability allows it to detect criminal activities swiftly and trigger emergency alerts, significantly reducing response times.

The combination of spatial (YOLOv11) and temporal (LSTM) analysis ensures accurate detection of both static and dynamic threats. The experimental results validate that the system can operate effectively in complex environments, demonstrating high precision and stability. This innovation contributes to the development of **AI-driven public safety systems** capable of assisting law enforcement and preventing crimes proactively.

7. FUTURE SCOPE

Although the system demonstrates robust performance, there are several avenues for further enhancement:

1. **Integration with IoT devices:**
   Connecting the system with smart sensors (motion, sound, or door alarms) for more comprehensive surveillance.
2. **Cloud-based scalability:**
   Deploying the system on cloud platforms for large-scale, city-wide monitoring and centralized control.

3. **Multimodal threat detection:**
   Including **audio analysis** (e.g., detecting screams or gunshots) to complement visual detection.

4. **Facial recognition module:**
   Integrating face recognition to identify known criminals or suspicious individuals.

5. **Edge AI implementation:**
   Deploying lightweight models on **Raspberry Pi or Jetson Nano** for on-site, low-latency processing.

6. **Predictive crime analytics:**
   Using historical data and behaviour analysis to predict and prevent potential crimes before they occur.

8. REFERENCES

[1] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NeurIPS. (classic two-stage detector).

[2] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. ECCV. (single-shot detector balancing speed & accuracy).

[3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection (YOLOv1). CVPR. (introduced single-pass real-time detection).

[4] Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. arXiv.(improvements in backbone and multi-scale prediction).

[5] Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv. (practical speed/accuracy improvements and real-time features).

[6] Jocher, et al. (2020). YOLOv5. Ultralytics (open-source implementation widely used in practice). (practical, easy-to-use family—numerous community variants).

[7] Ultralytics (2023). YOLOv8: The Next-Generation YOLO Family. Ultralytics docs & repo. (modernized training & export pipeline commonly used for edge/real-time).

[8] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection (RetinaNet). ICCV. (addresses class imbalance for dense detectors).

[9] Tan, M., Pang, R., & Le, Q. (2020). EfficientDet: Scalable and Efficient Object Detection. CVPR. (compound scaling for detection backbones and biFPN).

[10] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). DETR: End-to-End Object Detection with Transformers. ECCV. (reformulates detection as set prediction with transformers).

[11] Sultani, W., Chen, C., & Shah, M. (2018). Real-World Anomaly Detection in Surveillance Videos. CVPR. (UCF-Crime dataset; MIL-based weakly supervised anomaly detection).

[12] Cheng, M., Cai, K., & Li, M. (2019). RWF-2000: An Open Large Scale Video Database for Violence Detection. arXiv / RWF-2000 dataset paper. (real surveillance clips; Flow-Gated Network using optical flow + 3D CNN).

[13] Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., & Yang, Z. (2020). Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision (XD-Violence). ECCV. (XD-Violence dataset with audio+video; multimodal weakly-supervised approach).

[14] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast Networks for Video Recognition. ICCV. (powerful spatio-temporal models that are widely used for action and violence recognition).

[15] [Survey/Recent applied work] Efficient / Multimodal Violence Detection & Fusion approaches — examples include MDPI journal articles and applied works (2021–2024) that explore lightweight architectures, optical-flow fusion, and gated multimodal fusion for violence/anomaly detection (see representative articles in Sensors/Entropy).