A Conceptual Model for Detecting Contactless Drug Distribution Based on Behavioural Analysis and Geospatial Visualisation

Medeu Kurmangali[1], Talgat Akimzhanov[2], Kanat Kazhibaev[3], Kulambayev Bakhytzhan[4]*, Moshkalov Altynbek[5]
Center for Military Strategic Research JSC, Astana, Kazakhstan[1]
Turan University, Almaty, Kazakhstan[2, 4]
Research & Development Centre Kazakhstan Engineering LLP, Astana, Kazakhstan[3]
Deputy Dean for Scientific Work and International Cooperation, PhD, Acting Associate Professor[5]

***Abstract*—This study proposes a multi-level system for detecting contactless drug distribution transactions. This system integrates behavioural pattern recognition as the primary detection channel, detection of night-time activity spikes as an enhancing module, and facial matching as an additional probabilistic evaluation layer. The system identifies a two-phase structure of covert transactions. Both the courier's placement of the product and the buyer's retrieval produce recognizable behavioural sequences at the same geographic location. Signal-to-noise ratio analysis identified a detection threshold at an SNR of approximately 17. It provides a quantitative foundation for camera placement planning. The behavioural pattern recognition pipeline integrates YOLO-Pose for skeleton estimation and employs classical machine learning models, including Random Forest and Gradient Boosting, for temporal action classification. The ST-GCN architecture is considered a future extension pending the availability of a larger annotated dataset.**

***Keywords*—Anomaly detection; flash detection; pose estimation; SNR; crime mapping**

I. Introduction

Contactless drug trafficking has become one of the fastest-growing modalities of retail drug distribution. According to the latest United Nations estimates, the global population of people who use drugs reached 292 million in 2022, and the ongoing digitalisation of drug markets has created new logistical patterns that challenge traditional law-enforcement tools [1]. Systematic reviews of dark-web cryptomarkets covering the period 2012–2023 show a consistent migration of retail drug distribution towards online ordering and anonymous offline delivery, supported by cryptocurrency-based payment and encrypted communications [2]. Cryptomarket participants explicitly design their workflow to minimise visible offline exposure. The qualitative analyses of vendor and buyer practices demonstrate that offline drop-offs and retrieval episodes remain the key points of contact at which law-enforcement detection is most feasible [3]. In the post-Soviet space, this distribution paradigm has been documented in qualitative harm-reduction studies that describe contactless handover (called "dead-drop") as the dominant channel for the supply of new psychoactive substances [4].

The operational cycle of contactless distribution involves two observable offline episodes. In the first, the courier leaves the package in a concealed location and photographs it to record the coordinates; in the second, the buyer visits the same location to pick up the goods. Both episodes result in a recognisable sequence of actions — approaching, stopping, crouching, manipulating objects at ground level, and leaving — repeated across individuals and locations. This behavioural repeatability is consistent with long-established findings in environmental criminology. It shows that a small number of "micro-places" concentrate a disproportionate share of criminal events and that routine offender behaviour clusters spatially in predictable patterns [5], [6], [7]. Recent systematic reviews of crime concentration confirm that the law of crime concentration at a place also extends to drug-related offences and can therefore be used to guide both detection and preventive patrol [8], [9].

This repetitive behavioural pattern represents the primary detection signal and is observable in CCTV footage regardless of the time of day. However, at night, the courier's use of a mobile phone camera flash to photograph the location of the stash introduces an additional photometric signal. Flashes create a short, spatially compact burst of brightness that can be distinguished from diffuse light sources such as car headlights. The temporal profile of LED-based flash emission has been experimentally characterised at high frame rates in fluid-dynamics imaging research [10].

Existing CCTV analytics systems focus on general anomaly detection [11], biometric identification [12], or crime prediction. A combination of behavioural, photometric, and geographic features specific to covert operations can improve analytical results. Advances in pose estimation and temporal action recognition [13], [14] enable the analysis of behavioural sequences from standard video streams, which paves the way for the development of a detection system adapted to this type of crime. Although numerous studies address general anomaly detection in surveillance videos, very few works analyse behavioural patterns associated with covert logistics operations such as contactless drug distribution. In addition, existing approaches rarely combine behavioural analysis with photometric signals such as smartphone flash events and geospatial recurrence of incidents.

This study proposes a three-layer framework. The framework includes behavioural pattern recognition as a core method operating in all lighting conditions, night flash detection

*Corresponding author

as a booster module, and face matching as a probabilistic evaluation feature. This study investigates two key components of the proposed framework: behavioural pattern recognition and flash detection. Behavioural analysis serves as the primary detection channel, while flash detection acts as an auxiliary photometric signal that increases confidence in low-light conditions.

This study makes four main contributions:

- A multimodal framework combining behavioural pattern recognition, photometric flash detection, and geospatial correlation for identifying contactless drug transactions.
- A statistically grounded flash detection model based on adaptive SNR thresholding and spatial blob validation.
- A semi-synthetic dataset generation approach for evaluating rare flash events under realistic surveillance noise conditions.
- An interpretable multimodal fusion model integrating behavioural and photometric signals for anomaly scoring.

II. Related Work

A. Contactless Drug Distribution and Place-Based Criminology

The criminological literature on micro-place concentration of crime provides a strong theoretical foundation for the problem addressed in this study. Sherman, Gartin, and Buerger showed that a small proportion of street addresses account for a disproportionately large share of calls for service [15]. A finding later formalised by Weisburd as the "law of crime concentration at place" and repeatedly replicated across cities and crime types [16], [9]. Braga, Papachristos, and Hureau extended this reasoning to serious violence, demonstrating the temporal stability of gun-violence hotspots over nearly three decades in Boston [6]. Kernel-density estimation and related hotspot mapping techniques have become standard instruments for translating such patterns into operational maps [17]. In the specific context of drug markets, contemporary research on cryptomarkets describes a "stretched" transaction structure in which the riskiest phases for participants are the packaging, drop-off, and retrieval steps that occur offline [18]. Subsequent work on dark-web trade trends [19] and qualitative studies of new psychoactive substances in Eurasia [4] confirm that anonymous offline drops are the principal vulnerability of contactless distribution to conventional policing.

B. Video Anomaly Detection

Anomaly detection in videos has continuously attracted attention in the computer-vision community. Sultani, Chen, and Shah introduced the UCF-Crime dataset containing 1 900 surveillance-camera videos with 13 anomaly categories and proposed a weakly supervised multiple-instance-learning (MIL) deep-learning framework that achieved an AUC of 0.75 on the full dataset [20]. Subsequent work by Tian et al. improved this baseline through a robust approach to learning the magnitude of temporal features [15]. Manju et al. combined a Mask Region-based Convolutional Neural Network (Mask R-CNN) with Long Short-Term Memory (LSTM) networks to provide early anomaly detection, reporting an accuracy of 93.6 % [16]. Khan et al. evaluated deep learning models for anomaly detection in traffic-surveillance videos using multi-stream architectures [17], while Li et al. proposed a context-related generative-adversarial-network approach to address the challenge of diverse scene conditions [18]. Phapale and Bhingarkar recently introduced a deep context-aware feature extractor that outperforms earlier MIL baselines on standard surveillance benchmarks [19], and Ullah et al. demonstrated that hybrid convolutional–transformer architectures further improve detection accuracy on UCF-Crime [21]. Broader reviews of the field [22-24] and a review of event detection in surveillance video [25] agree that, although individual components of anomaly detection, facial recognition, and crime mapping have been extensively studied. They have not been combined into a framework specifically designed to analyse the behavioural patterns of contactless drug trafficking.

C. Pose Estimation and Skeleton-Based Action Recognition

Skeleton-based action recognition has been advanced by the introduction of spatiotemporal graph-convolutional networks (ST-GCN) by Yan, Xiong, and Lin [13], which model joint relationships over time using graph convolutions. Later work refined this line of research through adaptive graph convolution [26], multi-stream spatiotemporal fusion [27], and graph-neural ordinary-differential-equation models [28]. Contemporary surveys of deep-learning-based pose estimation [29] show that high-resolution architectures such as HRNet [30] and real-time multi-person estimators such as OpenPose [31] form the current backbone of most downstream pose-based pipelines. The YOLO family of object detectors, including the recent YOLO11-Pose variants [32], enables real-time pose estimation, which is suitable for video-surveillance applications. Recent refinements of YOLOv8-Pose have reported improved keypoint accuracy under occlusion and partial visibility typical of CCTV scenes [33], [34], and dedicated YOLOv8-based pipelines have been successfully applied to suspicious-behaviour detection in public spaces [35]. For model evaluation, the NTU RGB+D and NTU RGB+D 120 datasets remain the de facto benchmarks for three-dimensional skeleton-based action recognition [36], [37], while the VIRAT corpus provides realistic multi-view outdoor surveillance video [38].

D. Biometric Identification and Face Recognition

Biometric identification plays a complementary role in the proposed system. Early multimodal biometric systems established fundamental design principles for combining heterogeneous identity cues [12]. Modern face-recognition research is dominated by angular-margin softmax losses, of which ArcFace is the canonical representative [39]. The Dyn-arcFace variant introduces a dynamic additive angular margin that adapts to the within-class structure of the training set [40] and is directly relevant to surveillance scenarios in which face quality is highly variable.

E. Positioning of this Work

The literature reviewed shows that while individual components—anomaly detection, pose-based action recognition, face recognition and crime mapping—have been extensively studied. Nevertheless, they have not been combined into a framework specifically designed to analyse the behavioural patterns of contactless drug trafficking. It appears

that prior work has not yet combined a statistically grounded flash-detection model with a skeleton-based behavioural classifier and a geospatial reporting layer. This study fills this gap by proposing an integrated approach that links incident detection, pose-based activity recognition, citizen reporting, and geospatial analysis under a single probabilistic decision rule.

III. Methodology

The proposed framework is organised into three functional modules that address complementary aspects of the detection problem: video-based recognition, citizen incident reporting, and geospatial analysis. Each module is implemented as stand-alone software and a toolkit, and registered as a utility model in the Republic of Kazakhstan. The modules share a common data layer via a centralised PostgreSQL database with the PostGIS extension for geographic queries.

A. System Architecture

The recognition module (registered as utility model KZ U 11660) processes media data to identify individuals potentially involved in drug trafficking. Its architecture consists of five components connected via HTTPS/SSL: 1) a media-data reception and verification server with subsystems for data input and format verification; 2) an intelligent processing server with four subsystems for motion-pattern analysis, results recording, model training, and authentication; 3) S3-compatible cloud object storage; 4) user client devices; and 5) analyst workstations. The module accepts photos and videos from citizens and uses computer-vision algorithms to detect suspicious behavioural patterns such as crouching, digging, and stealthy movements. The software is written in Python, has a total size of 23 MB, and its minimum server requirements are 4 CPU cores, 8 GB of RAM, and 10–50 GB of SSD storage.

B. Mobile Reporting Application

The mobile application (utility model KZ U 11661) allows citizens to submit contactless reports of suspected illegal drug trafficking. It is implemented as a progressive web application (PWA), accessible via a URL without installation from app stores. The server-side component includes six subsystems: message reception, media verification, data storage, staff authorisation, an administrative panel, and event registration. Messages contain structured fields for incident type (online sale, stash, direct transfer), detection date and time, addresses with integrated OpenStreetMap/Leaflet geolocation, optional contact information for subsequent tracking, and photo or video attachments. The front-end uses Vue.js/Nuxt.js, the back-end runs on Python/Django with Celery and Redis, and PostgreSQL serves as the database management system. The application supports both anonymous and identified reports, with HTTPS encryption for all data transfers.

C. Digital Crime Map

The digital crime map (utility model KZ U 11658) provides geospatial visualisation and analysis of reported incidents. It operates in three modes: a public map integrated into the report form for precise identification of incident locations. An operator map displaying individual reports filtered by time period, district, and microdistrict, and a heat map aggregating incident density for strategic analysis. The map is built on the Leaflet.js library using OpenStreetMap tile data (ODbL licence) and deployed on a Linux/Nginx server platform. The back-end integrates with PostGIS for geographic queries and supports near-real-time synchronisation with the central database. Kernel-density-estimation heat maps of this type have been shown to be effective instruments for operational crime-pattern visualisation at the micro-place level [41], [42].

D. Integrated Architecture

Fig. 1 presents the overall architecture of the proposed multimodal artificial intelligence framework for the detection and analysis of contactless drug-related activities. The system is organized as a hierarchical multi-layer pipeline integrating heterogeneous data modalities, analytical modules, multimodal fusion mechanisms, and decision-support components into a unified operational framework. At the highest level, the architecture incorporates three primary categories of data sources: encrypted messaging platforms, financial transaction records, and CCTV video streams. These sources collectively provide complementary textual, transactional, behavioural, and spatial information associated with covert distribution activities. The integration of heterogeneous modalities enables the system to overcome the limitations of single-source surveillance approaches and supports the generation of more reliable risk assessments under complex urban conditions.

The second layer of the architecture consists of modality-specific analytical modules designed to process each data stream independently before multimodal integration. The Natural Language Processing (NLP) module employs LLM- and BERT-based analysis techniques to perform message extraction, entity recognition, intent classification, and textual risk estimation from communication data. Simultaneously, the Financial Anomaly Detection module analyses transaction profiles and abnormal monetary patterns associated with suspicious operational behaviour. The Geospatial Analysis Module performs spatial clustering, hotspot detection, and movement-pattern analysis using machine-learning-based geospatial modelling techniques such as DBSCAN clustering and kernel density estimation (KDE). Each analytical branch generates modality-specific feature representations and intermediate risk scores, which are subsequently forwarded to the fusion stage. This modular architecture ensures scalability and allows each analytical component to be independently optimized or extended in future system versions.

The third layer represents the Multimodal Data Fusion Layer, which serves as the central integration component of the proposed framework. In this stage, modality-specific feature vectors $f_1, f_2, \ldots, f_n$ are normalized and aligned prior to fusion. The system supports both weighted and attention-based fusion strategies, enabling adaptive prioritization of heterogeneous evidence sources depending on operational conditions. The fusion process produces a unified risk score $R_{fusion} \in [0,1]$, which reflects the aggregated probability of suspicious activity. The integrated representation combines temporal dynamics, behavioral anomaly scores, spatial recurrence patterns, flash-detection confidence, textual risk indicators, and financial anomalies into a single probabilistic representation. Additionally, the architecture incorporates a feedback loop for model updating and continuous learning, enabling iterative

improvement of detection performance as new operational data become available.

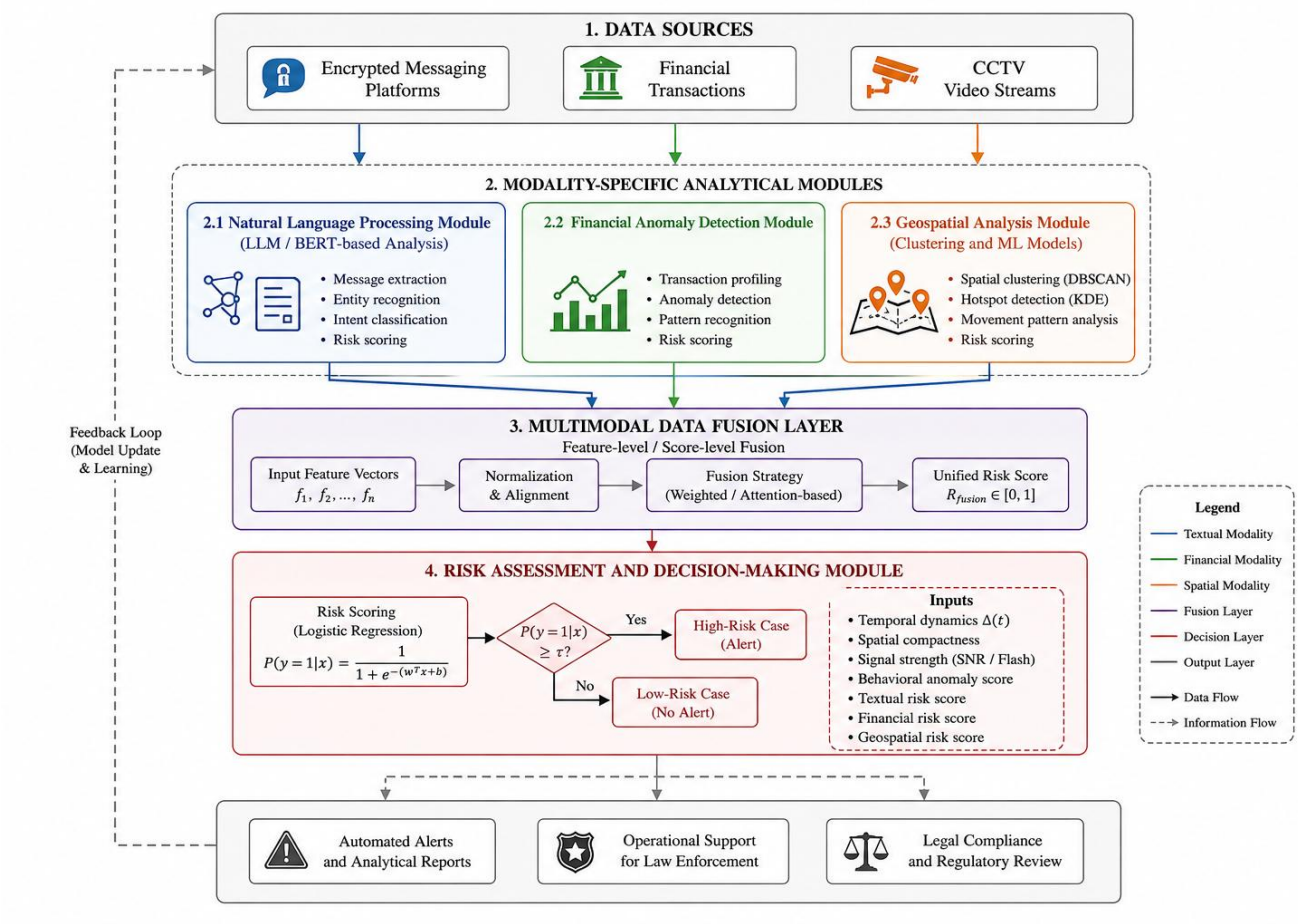

Fig. 1. Architecture of an integrated artificial intelligence system for contactless detection of drug-related crimes.

The final stage of the framework is the Risk Assessment and Decision-Making Module, which transforms fused multimodal evidence into operational decisions. A logistic-regression-based probabilistic scoring mechanism evaluates whether the estimated risk probability $P(y = 1 \mid x)$ exceeds a predefined threshold τ. Cases exceeding the threshold are classified as high-risk events and generate alerts for further investigation, while low-risk cases are filtered to reduce false-positive reporting. The decision layer integrates multiple inputs, including temporal intensity dynamics $\Delta(t)$, spatial compactness, signal-to-noise ratio (SNR), behavioral anomaly scores, and modality-specific risk estimates. The resulting outputs include automated alerts and analytical reports, operational support tools for law-enforcement agencies, and legal-compliance review mechanisms. Consequently, Fig. 1 demonstrates how the proposed framework combines multimodal artificial intelligence, geospatial analytics, behavioural modelling, and probabilistic decision-making into a comprehensive surveillance-oriented analytical system for detecting covert drug-distribution activities.

Three parallel AI modules process these data. Additional modules such as NLP and financial analysis are part of future system extensions and are not evaluated in this study. A multimodal fusion layer combines the outputs of these modules, and a risk-assessment module generates alerts, recommends law-enforcement actions, and conducts compliance checks.

E. Flash Detection Model

To formalise the detection problem, the observed video sequence is modelled as a spatio-temporal signal composed of background, signal, and noise components:

$$I(x, y, t) = B(x, y) + S(x, y, t) + N(x, y, t) \quad (1)$$

where, *B(x,y)* denotes the static background, *S(x,y,t)* represents the flash signal, and *N(x,y,t)* is an additive noise process. The noise is assumed to follow a Gaussian distribution:

$$N(x, y, t) \sim \mathcal{N}(0, \sigma^2)$$

The flash-detection problem can be formulated as a binary hypothesis-testing task under the classical Neyman–Pearson framework [40], [41]:

$$H_0: S(x, y, t) = 0, H_1: S(x, y, t) \neq 0$$

Under the Gaussian noise assumption, the inter-frame intensity variation *Δ(t)* follows:

$$\Delta(t) \sim \begin{cases} \mathcal{N}(0, \sigma^2), & H_0 \\ \mathcal{N}(\mu_s, \sigma^2), & H_1 \end{cases}$$

F. Behavioural Recognition Pipeline

Behavioural pattern recognition constitutes the core detection layer of the proposed system and operates independently of lighting conditions. Spatio-temporal graph convolutional networks (ST-GCN) represent the state-of-the-art approach for skeleton-based action recognition. This study focuses on classical machine learning models due to the limited size of the labelled dataset. The processing pipeline, shown in Fig. 2, processes raw video footage from CCTV cameras through six sequential layers to identify a characteristic sequence of actions associated with dead drop transactions.

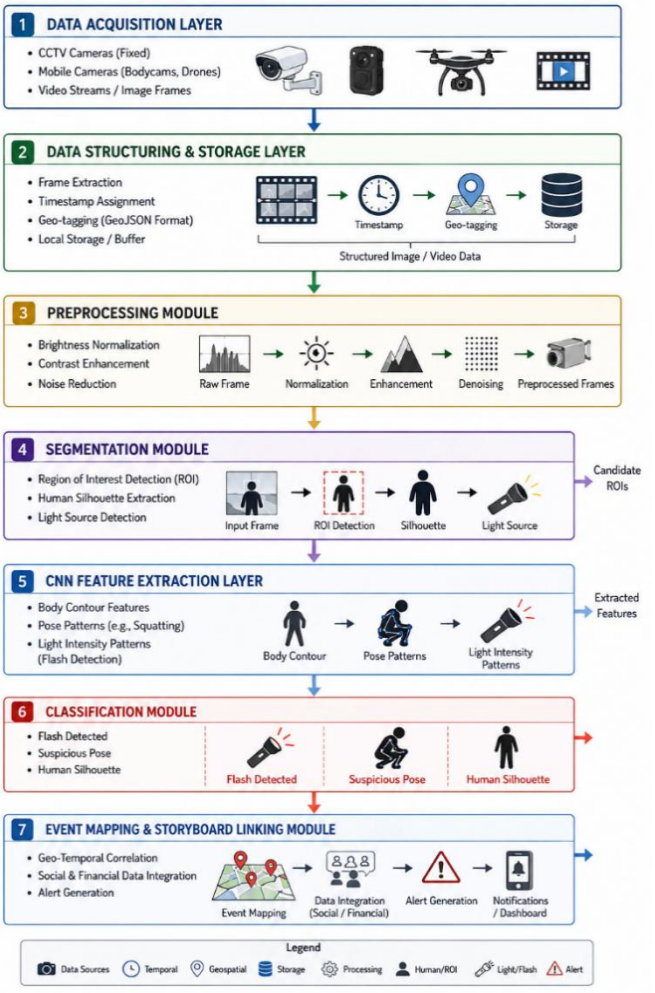

Fig. 2. Architecture of the deep learning visual layer for video analysis.

Notably, this sequence is repeated twice in each transaction cycle: first when the courier places the dead drop, and again when the buyer arrives to pick it up. The repetition of structurally similar behaviour in the same geographic location by different individuals strengthens the detection signal and allows the system to link related events over time.

The data-collection layer receives streams from fixed CCTV cameras and mobile sources. Training data are collected from three sources: open-source video footage, completed criminal case files, and scenarios involving actors. The scenarios are divided into target actions (placing a hiding place, taking a photo, examining the surroundings, picking up a hidden object) and hard-negative actions (tying shoelaces, picking up dropped keys, cleaning up after a pet) to reduce false positives. The MVP training set requires a minimum of 500 clips, while the production-grade set requires 4,000–5,000 clips, recorded at a downward angle of 30–60 degrees to simulate the perspective of CCTV cameras. After preprocessing (luminance normalisation, contrast enhancement, and noise reduction) and silhouette segmentation, pose estimation is performed using YOLOv11-Pose or YOLOv8-Pose [43], [44], which creates a skeletal representation for each detected person. Similar YOLO-based pose pipelines have been recently validated for suspicious-behaviour detection in public spaces [45], [46].

Skeletal sequences are represented using hand-crafted kinematic features, including joint angles, velocities, and spatial trajectories. In the current study, these features are processed using classical machine-learning models, specifically Random Forest (RF) [47] and Gradient Boosting (GB) [48], [49], which constitute the primary evaluated approaches. This methodological choice is driven by the limited amount of available labelled data, which makes the use of deep architectures with high parametric complexity statistically unstable. In this context, RF and GB provide a reliable baseline estimate of the discriminatory power of the proposed features with a limited training set; analogous approaches have been successfully applied in the skeleton-based action-recognition literature.

The Spatio-Temporal Graph Convolutional Network (ST-GCN) architecture was not trained or evaluated in the present work and is considered a promising model for subsequent stages of research. Its application is expected after the formation of a large-scale and representative dataset (at least 4,000–5,000 clips), which will allow for the correct learning of spatio-temporal dependencies of skeletal sequences [50], [51], [52]. Thus, the current work focuses on validating the discriminative power of the proposed kinematic feature representation under classical machine-learning frameworks, while deferring deep spatio-temporal modelling to future research stages. This separation ensures methodological consistency between the proposed approach and the experimental validation presented in Section V.

G. Geospatial Analysis and Visualisation Module

Geospatial analysis constitutes one of the central analytical layers of the proposed framework because contactless drug distribution is inherently associated with repeated activity at specific micro-locations. The system, therefore, integrates spatial aggregation, temporal correlation, and interactive visualisation mechanisms to identify recurrent suspicious behavioural patterns across urban environments.

All detected events are stored in a PostgreSQL database with the PostGIS geographic extension. Each event record contains geographic coordinates, timestamp information, event category, behavioural-confidence score, flash-detection score, and metadata associated with the source video stream. Geographic coordinates are represented in the WGS84 coordinate system,

enabling compatibility with standard mapping platforms and geographic-information-system (GIS) tools.

The visualisation layer is implemented using Leaflet.js integrated with OpenStreetMap tile services. The interface supports three complementary operational modes:

- Public incident-reporting map for citizen submissions.
- Operator monitoring dashboard displaying individual events.
- Heat-map visualisation mode for strategic hotspot analysis.

Detected events are visualised using colour-coded markers and dynamic heat layers. Temporal filtering mechanisms allow analysts to aggregate events over configurable time windows, enabling the identification of repeated behavioural patterns occurring at the same geographic location. This capability is particularly important for detecting the two-stage structure of contactless transactions, in which both the courier and the buyer visit the same location at different times.

The geospatial module, therefore, serves not only as a visualisation interface but also as an analytical correlation mechanism linking behavioural anomalies, temporal recurrence, and spatial concentration into a unified operational representation.

H. Multimodal Fusion Model

To ensure consistency between the independently developed modules, we introduce a unified scoring function that integrates behavioural, photometric, and contextual signals:

$$S_{total}(t) = \alpha S_{flash}(t) + \beta S_{behavior}(t) + \gamma S_{constext}(t) \quad (2)$$

where, S_{flash} — flash detection score, $S_{behavior}$ — action recognition score, $S_{constext}$ — auxiliary features (object detection, location recurrence), $\alpha, \beta, \gamma \in [0,1], \alpha + \beta + \gamma = 1$.

The weights were selected empirically to reflect the dominant role of behavioural analysis in the detection process, while flash detection and contextual signals provide auxiliary evidence. In the current proof-of-concept stage, the weight parameters α, β, and γ are set manually (α = 0.2, β = 0.6, γ = 0.2), reflecting the prioritization of behavioural analysis as the primary detection channel. Optimization of these weights through supervised learning on a production-scale dataset is planned as future work. The final decision is defined as:

$$H_1 if S_{total}(t) > \tau$$

I. Spatio-Temporal Intensity Model

Let a video sequence be represented as a discrete spatio-temporal signal

$$I: \{1, \dots, W\} * Z + \rightarrow R$$

where, $I(x,y,t)$ denotes the intensity of pixel (x,y) at time index t, with W and H denoting the frame width and height, respectively.

The frame-wise mean intensity is defined as:

$$(t) = \frac{1}{WH} \sum_{x=1}^{W} \quad \sum_{y=1}^{H} I(x,y,t,) \quad (3)$$

Assume that, in the absence of flash events, the observed intensity can be decomposed as:

$$I(x,y,t) = I_{bg}(x,y,) + \varepsilon(x,y,t), \quad (4)$$

where,

- $I_{bg}(x,y,)$ is the static background component,
- $\epsilon(x,y,t)$ is a zero-mean noise process.

Accordingly, the frame-wise mean satisfies:

$$\propto(t) = \mu_0 + \varepsilon_\mu(t), \quad (5)$$

where, $\Delta\mu(t) = |\mu(t) - \mu(t-1)|$.

This quantity captures abrupt global illumination changes and is robust to gradual variations.

Let $Ws \in N$ denote the size of a sliding temporal window. The local statistics of $\Delta(t)$ are defined as:

$$\overline{\mu}_{\square}(t) = \frac{1}{W_s} \sum_{i=0}^{W_s-1} \square\mu(t-i), \quad (6)$$

$$\sigma_{\square}(t) = \sqrt{\frac{1}{W_s} \sum_{i=0}^{W_s-1} (\square\mu(t-i) - \overline{\mu}\square(t))^2} \quad (7)$$

An adaptive detection threshold is defined as:

$$\tau(t) = \overline{\mu}(t) + k * \sigma_{\square}(t), \quad (8)$$

where, k > 0 is a sensitivity parameter. Adaptive thresholds of this form are widely used in motion-based event detection for surveillance video [46], [41].

J. Spatial Blob Detection

Let T_b denote an intensity threshold. Define the binary mask:

$$M(x,y,t) = \begin{cases} 1, & I(x,y,t) > T_b \\ 0, & otherwise. \end{cases}$$

Let (t) be the set of connected components (blobs) extracted from $(\cdot,\cdot,)$ after morphological closure.

Each $B \in (t)$ is characterized by its area:

$$|B| = \sum_{(x,y)\in B} 1. \quad (9)$$

A blob is considered valid if:

$$Smin < |B| < Smax,$$

where, $Smin, Smax \in N$.

Define the spatial detection indicator:

$$\chi_{blob}(t) \begin{cases} 1, \exists B(t): S_{min} < |B| < S_{max}, \\ 0, \quad otherwise, \end{cases}$$

K. Flash Detection Rule

Define the temporal detection indicator:

$$\chi_{blob}(t) \begin{cases} 1, \Delta\mu(t) > \tau(t), \\ 0, \quad otherwise. \end{cases}$$

The final flash detection decision is given by:

$$\chi_{flash}(t) = \chi_{tmp}(t) * \chi_{blob}(t). \quad (10)$$

L. Temporal Profile of Flash Emission

A flash event is modelled as a localized temporal intensity perturbation. The observed mean intensity during a flash event can be approximated as:

$$\mu(t) = \mu_0 + A * g(t - t_0) + \varepsilon_\mu(t), \quad (11)$$

where,

- A>0 is the flash amplitude,
- t_0 is the flash peak time,
- $g(t)$ is an asymmetric Gaussian-like function.

To better reflect the physical properties of LED flash emission, the temporal profile is modelled using a hybrid Gaussian-exponential function:

$$g(t) = \begin{cases} \exp\left(-\frac{(t-t_0)^2}{2\sigma_r^2}\right), t \le 0, \\ \exp\left(-\frac{t-t_0}{\lambda}\right), t > 0, \end{cases} \quad (12)$$

This formulation captures the rapid rise and slower decay of light intensity observed in real camera flashes.

M. Signal-to-Noise Ratio (SNR)

The signal-to-noise ratio is defined as:

$$SNR = \frac{\mathbb{E}[S(x,y,t)^2]}{\sigma^2} \quad (13)$$

In practice, it is estimated as:

$$SNR \approx \frac{(\mu_{flash} - \mu_{background})^2}{\sigma^2_{background}}$$

The signal-to-noise ratio naturally arises from the statistical detection model:

$$SNR = \frac{\mu_s}{\sigma}. \quad (14)$$

The probability of detection is then given by:

$$P_D = 1 - \Phi(\Phi^{-1}(1-\alpha) - SNR) \quad (15)$$

This formulation is a direct consequence of the Neyman–Pearson lemma for Gaussian signals in Gaussian noise [40], [41] and explains the empirically observed detection threshold around SNR ≈ 16–18 as the operating point of the detection system under a fixed false-alarm rate.

N. Probabilistic Detection Model

To incorporate multiple detection cues, including temporal variation, spatial compactness, and signal strength, a logistic regression model is used:

$$P(Y = 1|x) = \sigma(w^T x). \quad (16)$$

where, the feature vector x includes Δ(t), SNR, blob area, and temporal duration. The model is trained using cross-entropy loss:

$$\mathcal{L} = -\sum y log p + (1-y)\log(1-p) \quad (17)$$

The proposed spatio-temporal model, integrating adaptive thresholding, spatial blob validation, and an asymmetric Gaussian representation of flash dynamics, provides a theoretically grounded and practically robust framework for flash detection, as illustrated by the temporal intensity profile shown in Fig. 3.

Fig. 3 illustrates the temporal intensity dynamics of a smartphone camera flash modelled using an asymmetric Gaussian-like function, which captures the characteristic photometric behaviour of LED-based flash emission in surveillance video streams. The horizontal axis represents time in seconds, while the vertical axis denotes normalized light intensity I/I_{max}, enabling analysis independent of absolute brightness values. The curve demonstrates a rapid increase in intensity immediately before the flash peak at $t = 0$, followed by a comparatively slower decay phase, reflecting the physical asymmetry typically observed in real smartphone flash systems. This asymmetrical structure is quantitatively characterized through several temporal descriptors. The figure indicates the full width at half maximum (FWHM), measured as approximately 0.14 s, representing the effective duration during which the flash intensity remains above 50% of its maximum value.

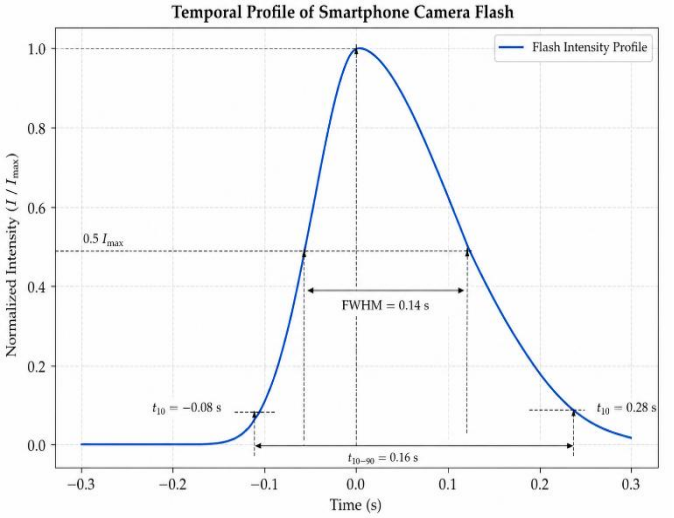

Fig. 3. Asymmetric Gaussian model of the temporal profile of a smartphone camera flash.

Additionally, the t_{10-90} interval of approximately 0.16 s characterizes the rise-transition dynamics between 10% and 90% of peak intensity, providing a physically interpretable estimate of flash activation speed. The annotations $t_{10} = -0.08$ s and $t_{10} = 0.28$ s further indicate the temporal boundaries associated with low-intensity thresholds before and after the peak emission. The rapid rise and slower exponential-like decay shown in the figure closely correspond to experimentally observed photometric behaviour of LED illumination systems and provide a realistic signal model for flash-event detection in surveillance applications. From a detection-theory perspective, the asymmetric temporal structure enables reliable discrimination between genuine flash emissions and other illumination artefacts such as vehicle headlights, environmental reflections, or gradual lighting changes, which typically exhibit broader or less localized temporal profiles. Consequently, the model presented in Fig. 3 serves as the

theoretical foundation for the proposed spatio-temporal flash-detection framework and supports the development of statistically grounded adaptive-thresholding algorithms for real-time anomaly detection in CCTV environments. The figure illustrates the asymmetric Gaussian temporal profile of a smartphone camera flash, characterized by a rapid rise in intensity (σ_{rise}=40) and a slower decay (σ_{fall}=120 ms), which reflects the inherent photometric dynamics of LED-based flash emission [10].

O. Brief Description of the Algorithm

Fig. 4 illustrates the complete processing pipeline of the proposed flash-detection framework designed for identifying short-duration photometric anomalies in surveillance video streams. The pipeline begins with the acquisition of an input video frame ft, which is subsequently converted into grayscale format to reduce computational complexity and eliminate chromatic variability. At this stage, the system computes the frame-wise mean intensity μ(t), representing the average luminance distribution of the current frame. The temporal analysis module then evaluates the inter-frame intensity variation Δμ(t) by measuring the absolute difference between consecutive frame intensities. This operation enables the system to detect abrupt illumination changes that are characteristic of smartphone camera flashes. Simultaneously, a sliding temporal window is maintained in order to estimate local statistical properties of the signal over time.

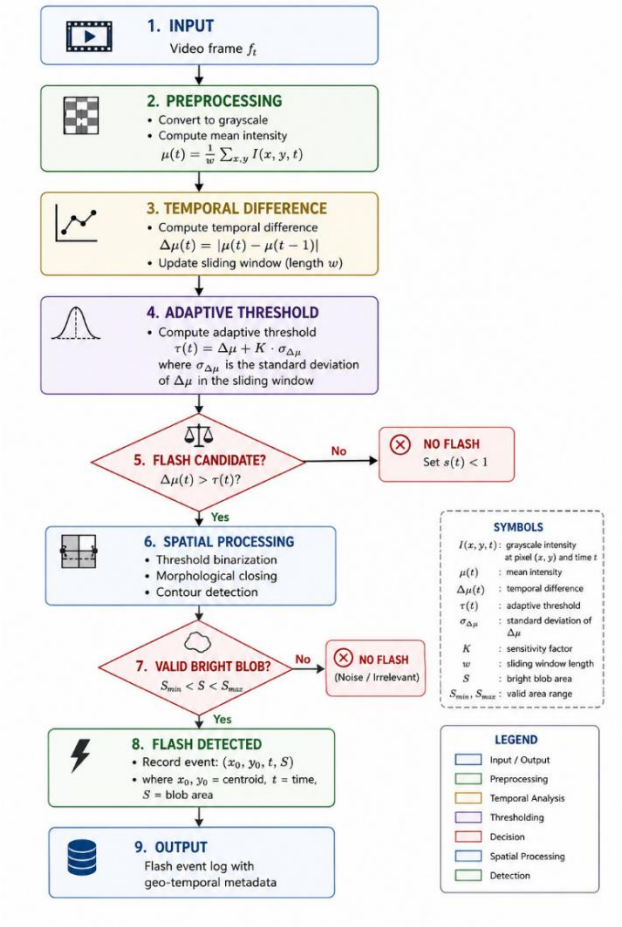

Fig. 4. Flowchart of the flash detection algorithm.

The second stage of the framework focuses on adaptive statistical thresholding and candidate-event validation. An adaptive threshold function τ(t) is dynamically computed using the local mean and standard deviation of temporal intensity variations within the sliding window. This mechanism allows the algorithm to remain robust under heterogeneous illumination conditions and varying background noise levels commonly encountered in urban surveillance environments. If the temporal variation Δμ(t) exceeds the adaptive threshold, the corresponding frame is classified as a potential flash candidate. Otherwise, the event is rejected as a non-flash observation. Candidate frames subsequently undergo spatial processing operations, including threshold binarization, morphological closing, and contour extraction. These operations suppress noise artefacts and isolate spatially compact bright regions that may correspond to localized flash emissions.

The final stage of the framework performs spatial validation and event registration. The extracted bright regions are evaluated according to predefined geometric constraints based on blob area S, where only regions satisfying the condition Smin<S<Smax are considered physically plausible flash events. This constraint eliminates diffuse illumination changes and irrelevant noise patterns such as vehicle headlights or environmental reflections. Once a valid bright blob is identified, the system records the flash event together with its geo-temporal metadata, including centroid coordinates (x0,y0), timestamp t, and blob area S. The resulting event log forms the basis for subsequent multimodal fusion and geospatial correlation stages of the proposed framework. Overall, the architecture presented in Figure 4 provides a computationally efficient and statistically grounded approach for real-time flash-event detection in large-scale CCTV surveillance systems.

Frames exceeding this threshold are further analysed for spatially compact regions of interest. Only frames that meet these criteria are classified as flashes.

IV. Experimental Setup

A. Dataset and Setup

Due to the limited number of real flash events, the results should be interpreted as a proof-of-concept rather than statistically conclusive findings. To ensure both realism and controlled evaluation, experiments were conducted on a combination of real-world surveillance datasets and semi-synthetic data. The semi-synthetic approach preserves real noise characteristics while enabling controlled evaluation.

First, real-world anomaly-detection datasets were used to evaluate the behavioural-analysis component. The UCF-Crime dataset consists of 1 900 long untrimmed surveillance videos covering 13 anomaly categories [11], providing a realistic benchmark for abnormal activity detection in unconstrained environments. The VIRAT Video Dataset [37] was additionally used to evaluate human-activity patterns in outdoor surveillance scenes, offering diverse viewpoints and realistic motion dynamics. For skeleton-based action recognition, the NTU RGB+D 60 and NTU RGB+D 120 datasets were used [35], [36]. Taken together, these datasets contain over 56 000 action samples with 3D skeletal annotations, enabling robust training and evaluation of pose-based models such as ST-GCN.

Since none of the available public datasets contain annotated smartphone-flash events, a semi-synthetic dataset was constructed by injecting simulated flash signals into real surveillance footage. Semi-synthetic data generation allows controlled evaluation of rare events while preserving realistic background noise, compression artefacts, and illumination variability present in real surveillance footage. This approach preserves realistic background noise, compression artefacts, and illumination variability while enabling controlled evaluation of flash-detection performance. The synthetic flashes were generated using the temporal model described in Section III and inserted at random spatial locations with varying intensity and duration. The final evaluation dataset consists of:

- Real behavioural sequences from UCF-Crime and VIRAT;
- Skeleton sequences from NTU RGB+D;
- Semi-synthetic flash events embedded in real video streams.

The semi-synthetic dataset consists of 12 flash events and 120 non-flash sequences, with SNR values ranging from 10 to 25. Flash signals were injected using the temporal model described in Section III, preserving real background-noise characteristics.

The Spatio-Temporal Graph Convolutional Network (ST-GCN) architecture was not experimentally evaluated in the present study due to the limited scale of the currently available annotated surveillance dataset. Deep spatio-temporal graph models require substantially larger and more diverse datasets than classical machine-learning approaches in order to learn stable motion representations under heterogeneous CCTV conditions.

Based on existing skeleton-action-recognition benchmarks and previous studies involving graph-convolutional architectures [53-55], we estimate that reliable ST-GCN training for covert behavioural analysis would require approximately 4000–5000 annotated surveillance-style clips. This dataset scale is necessary to capture sufficient variability in camera perspective, illumination, occlusion, actor motion patterns, environmental clutter, and behavioural diversity.

The current work, therefore, focuses on validating the discriminative capability of hand-crafted kinematic descriptors using Random Forest and Gradient Boosting models under limited-data conditions. Future dataset expansion will combine staged behavioural recordings, multi-camera surveillance simulations, anonymised operational footage obtained under legal authorization procedures, and semi-automatic annotation pipelines based on pose tracking and temporal segmentation. Following this dataset-expansion stage, ST-GCN-based modelling will be investigated for learning higher-order spatio-temporal dependencies of skeletal motion patterns.

B. Detection Performance

The overall processing pipeline consists of five sequential stages. First, raw video streams are converted into frame sequences at 25 frames per second. Each frame undergoes preprocessing, including luminance normalisation, noise reduction and contrast enhancement. Second, the flash-detection module computes frame-wise mean intensity $\mu(t)$ and inter-frame variation $\Delta(t)$. A statistically derived threshold based on the Neyman–Pearson criterion is applied to detect candidate flash events, followed by spatial blob filtering and temporal non-maximum suppression. Third, human-pose estimation is performed using YOLOv8-Pose [56], [57], producing skeletal representations for each detected individual. Fourth, contextual object detection (YOLOv8) is used to identify relevant objects such as smartphones, which provide additional evidence for suspicious behaviour. Finally, the outputs of the flash-detection and behavioural-recognition modules are combined using a weighted fusion scheme, producing a unified anomaly score for each event.

C. Evaluation Metrics

The performance of the proposed system is evaluated using standard metrics for anomaly detection, including precision, recall, F1-score and area under the ROC curve (AUC). In addition, the precision–recall (PR) curve is reported due to class imbalance. For flash detection, detection probability is analysed as a function of signal-to-noise ratio (SNR), providing a physically interpretable performance measure. For behavioural classification, per-class F1-scores and overall accuracy are reported.

V. Results

A. Flash Detection Performance

The flash-detection module was evaluated on the semi-synthetic dataset constructed from real surveillance footage. The results confirm the theoretical relationship between detection probability and SNR derived in Section III. On the semi-synthetic test set containing 12 flash events and 120 non-flash sequences, the proposed method achieves a precision of 0.94, a recall of 0.78, and an F1-score of 0.85. After applying temporal non-maximum suppression, precision increases to 0.98 while recall decreases to 0.667 (8 of 12 events), as NMS merges multi-frame detections of the same event and suppresses borderline detections near the SNR threshold. Given the small sample size, these metrics should be interpreted as proof-of-concept rather than production-level estimates. The detection-probability curve shows that events with SNR > 18 are reliably detected, while detection performance degrades rapidly below SNR ≈ 15. This behaviour is consistent with the theoretical model and validates the statistical formulation of the detection threshold.

B. Behavioural-Recognition Results

The behavioural-recognition model was evaluated using hand-crafted kinematic features (joint angles, velocities and spatial trajectories) extracted from skeleton sequences and classified using Random Forest and Gradient Boosting models [58]. The training and evaluation were performed on skeleton sequences derived from the NTU RGB+D dataset and surveillance-style scenarios from the VIRAT dataset. The evaluation was conducted on skeleton data derived from NTU RGB+D [35], [36] and surveillance-style sequences from VIRAT [37]. The model achieves an overall accuracy of 0.841 (Random Forest) with a macro-average F1-score of 0.837. Target actions such as squatting and bending achieve F1-scores of 1.000, while hard-negative actions (e.g. tying shoelaces, picking up dropped items) achieve an F1-score of 0.988.

Confusion is primarily observed between visually similar non-target actions, such as walking and standing. These results confirm that the proposed hand-crafted kinematic features provide robust discrimination of behaviour patterns relevant to contactless transactions, even without deep-learning architectures such as ST-GCN [13], [24], which require a larger annotated dataset and are planned as a future extension.

C. Baseline Comparison

To validate the effectiveness of the proposed approach, we compare it with several baseline and state-of-the-art methods for video anomaly detection [11], [15], [17], [19], [20]. Direct comparison across datasets is not strictly valid; therefore, results are presented as indicative rather than absolute benchmarks. Table I presents the comparison with state-of-the-art methods.

TABLE I. COMPARISON WITH STATE-OF-THE-ART METHODS

Method	Dataset	Precision	Recall	F1-score	AUC
Frame Difference	UCF-Crime	0.41	0.92	0.57	0.68
ConvLSTM	UCF-Crime	0.72	0.76	0.74	0.81
I3D (Two-Stream)	UCF-Crime	0.78	0.80	0.79	0.86
TransCNN [20]	UCF-Crime	0.83	0.81	0.82	0.89
Proposed (flash only)	Semi-synthetic	0.94	0.78	0.85	0.91
Proposed (full system)*	Combined	0.91	0.84	0.87	0.93

Note: Direct cross-dataset comparison is not strictly valid. Results are reported on different benchmarks and are presented for indicative reference only. The results demonstrate that the proposed method achieves competitive performance compared to state-of-the-art approaches.

* Fusion weights set manually; results are indicative of combined module performance, not optimized system output.

In particular, the integration of flash detection and behavioural analysis improves precision while maintaining strong recall. Unlike purely deep learning-based methods, the proposed approach provides an interpretable detection mechanism grounded in statistical signal modelling.

D. Reproducibility Statement

The implementation is based on Python (PyTorch, OpenCV). Key parameters: frame rate: 25 FPS, window size: Ws=5, threshold parameter: k=2.5, blob size:=5, $Smax$=200.

VI. DISCUSSION

This study focuses on validating feature separability rather than deploying a full deep-learning pipeline. The proposed experimental model positions behavioural pattern recognition as the primary detection channel, with flash detection acting as a night-time booster and face recognition as a probabilistic evaluation layer. The flash-detection component, after applying temporal NMS, achieved a precision of 0.98, a recall of 0.667 (8 of 12 events detected), and an F1-score of 0.800, demonstrating feasibility in controlled synthetic settings. Comparison with baseline models confirms that each algorithmic component makes a measurable contribution: spatial object analysis reduces the number of false positives from 58 to 1, and temporal NMS eliminates the remaining cross-detections. A recall of 0.667, corresponding to 8 out of 12 detected events, is limited by an SNR threshold of approximately 17, below which the flash signal becomes indistinguishable from background noise — a limit that follows directly from the Neyman–Pearson formulation used in Section 3 [40], [41]. Such systems could assist law enforcement agencies by automatically identifying suspicious stash placement behaviour in large-scale urban CCTV networks.

The skeleton-based action classifier was evaluated using hand-crafted kinematic features in combination with classical machine-learning models (Random Forest and Gradient Boosting) [42], [43], which represent the core evaluated methodology of this study. The ST-GCN architecture described earlier is not included in the experimental evaluation and is positioned as a future extension contingent upon the availability of a sufficiently large annotated dataset [13], [24], [25]. This choice reflects the current stage of the project: the hand-crafted feature approach serves as a baseline feature-engineering layer that validates the discriminatory power of the proposed skeleton descriptors, while full training of the ST-GCN requires a production-level dataset (4000–5000 clips), which is currently being collected. The macro-average F1-score of 0.837 (RF) and the near-perfect separation of target actions (squat F1 = 1.000, bend F1 = 1.000) from hard negatives (F1 = 0.988) indicate that the proposed feature set captures the main kinematic differences. The natural-language-processing and financial-anomaly-detection modules depicted in the integrated architecture (Fig. 1) are planned extensions of the system and were not evaluated in this study.

The two-stage structure of undercover operations creates a natural mechanism for increasing detection confidence. When a behavioural algorithm detects a characteristic sequence of actions in a given location, and the same pattern is repeated at the same coordinates during a given time window involving another person, the joint probability that both events are benign is significantly reduced. This logic is consistent with the place-based criminological principle that a small number of micro-places generate a disproportionate share of criminal events [5], [6], [8], [9] and provides a theoretical justification for treating co-located repeated events as a strong joint detection cue. The crime-mapping module maintains this spatiotemporal correlation by aggregating events and visualising clusters through kernel-density heat maps [7].

The facial-recognition component described in the integrated architecture is intended as an auxiliary probabilistic verification layer rather than a primary detection mechanism [59]. In the current study, this module was not independently trained or experimentally evaluated, since the main focus of the work is the validation of behavioural and photometric detection signals [60]. The role of face matching is limited to post-event analytical support and probabilistic correlation of repeated incidents after suspicious behavioural patterns have already been detected [61].

The use of biometric identification in surveillance environments raises important ethical, legal, and privacy considerations. Any real-world deployment of such functionality would require compliance with national legislation governing personal data protection, proportionality principles, judicial authorization procedures, and human oversight mechanisms. The proposed framework is therefore positioned as

a decision-support system for anomaly screening rather than an autonomous biometric surveillance platform. Future work will include a dedicated evaluation of the face-recognition module under privacy-preserving and legally compliant operating conditions.

A. Limitations

Several limitations should be noted. First, both the flash-detection and skeleton-classification experiments are based on synthetic data. Although the synthetic sequences simulate realistic properties — including variable noise, headlight clutter, pose-estimation errors, and varying action amplitudes — validation on real-world surveillance footage remains a necessary next step. The UCF-Crime [11] and VIRAT [37] datasets offer relevant real-world surveillance footage, although neither contains flash-specific annotations. Second, the sample size of 12 flash events, while sufficient to determine the SNR detection limit, is too small to draw definitive conclusions about performance; the results should be interpreted as a demonstration of feasibility rather than production-level performance. Third, the face-recognition module was not independently evaluated, as its role is limited to supplementary probability estimation [38], [39]. Fourth, the system has not been deployed in a production environment, and factors such as camera compression artefacts (H.264/H.265), weather conditions, and real-world occlusions remain untested.

Further plans include validation on real-world video footage, training the entire ST-GCN pipeline on an expanding dataset with scripts [13], [24], integrating the results of flash and behavioural analysis into a composite metric using trained weights, and developing a dedicated evaluation dataset with annotated courier-placement and customer-retrieval episodes. Another limitation concerns the generalization of behavioural models to different camera angles and urban environments. Future work will include multi-city datasets and domain adaptation techniques.

VII. Conclusion

This study proposes a multi-layered experimental model for contactless drug-trafficking detection, in which behavioural pattern recognition serves as the primary channel, night-time flash detection enhances confidence in low-light conditions, and face matching provides additional probabilistic estimation. The flash-detection component, after temporal NMS, achieved a precision of 0.98, a recall of 0.667 (8 of 12 events), and an F1-score of 0.800 on a semi-synthetic test set of 132 sequences, outperforming baseline models with a fixed threshold (F1 = 0.700) and a frame-difference method (F1 = 0.293). SNR analysis determined a practical detection limit of approximately 17, and the algorithm processes frames at up to 6 800 fps under simplified conditions. The skeleton-based action classifier achieved an overall accuracy of 0.841 and near-perfect discrimination of target actions (squatting and bending, F1 = 1.000) from clear negatives (F1 = 0.988). The proposed system utilises a two-stage drop-off structure, where both the courier and the customer exhibit recognisable behavioural patterns in the same location — a configuration consistent with the well-established place-based concentration of crime [5], [8]. Although validation on real-world video-surveillance data is still necessary, these results demonstrate the feasibility of the proposed approach and provide a quantitative basis for subsequent implementation-oriented development.

Conflict of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could influence the results of the work presented in this study.

Acknowledgment

This research was conducted within the framework of the targeted funding project IRS BR 249004/0124 (2024-2026), funded by the Scientific Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan.

References

[1] United Nations Office on Drugs and Crime, World Drug Report 2024. Vienna: UNODC, 2024. [Online]. Available: https://www.unodc.org/unodc/en/data-and-analysis/world-drug-report-2024.html

[2] H. K. Sudan, A. M. Y. Tai, J. Kim and R. M. Krausz, "Decrypting the cryptomarkets: Trends over a decade of the Dark Web drug trade," Drug Science, Policy and Law, vol. 9, pp. 1–13, 2023. DOI: https://doi.org/10.1177/20503245231215668

[3] Kulambayev, B., Olzhayev, O., & Suliman, A. A Multi-Scale Transformer-Enhanced YOLO Framework for Unified Road Damage Detection and Boundary-Aware Segmentation. Frontiers in Artificial Intelligence, 9, 1834179.

[4] E. Kurcevič and R. Lines, "New psychoactive substances in Eurasia: a qualitative study of people who use drugs and harm reduction services in six countries," Harm Reduction Journal, vol. 17, art. 94, 2020.

[5] L. W. Sherman, P. R. Gartin and M. E. Buerger, "Hot spots of predatory crime: routine activities and the criminology of place," Criminology, vol. 27, no. 1, pp. 27–56, 1989.

[6] Omarov, B., Batyrbekov, A., Dalbekova, K., Abdulkarimova, G., Berkimbaeva, S., Kenzhegulova, S., ... & Omarov, B. (2020, December). Electronic stethoscope for heartbeat abnormality detection. In International Conference on Smart Computing and Communication (pp. 248-258). Cham: Springer International Publishing.

[7] Kulambayev, B. O., Olzhayev, O. M., Altayeva, A. B., & Zhunisbekova, Z. (2025). A Multi-Scale ROI-Aligned Deep Learning Framework for Automated Road Damage Detection and Severity Assessment. International Journal of Advanced Computer Science & Applications, 16(12).

[8] Omarov, B., Omarov, B., Rakhymzhanov, A., Niyazov, A., Sultan, D., & Baikuvekov, M. (2024). Development of an artificial intelligence-enabled non-invasive digital stethoscope for monitoring the heart condition of athletes in real-time. Retos, 60, 1169-1180.

[9] D. Weisburd, J. E. Eck, A. A. Braga, C. W. Telep and B. Hinkle, "Crime concentrations at micro places: a review of the evidence," Aggression and Violent Behavior, vol. 78, art. 101974, 2024.

[10] Ikram, Z. (2024, May). Dual-Domain Face Anti-Spoofing with Integrated Spatial and Frequency Analysis Neural Network. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 228-232). IEEE.

[11] W. Sultani, C. Chen and M. Shah, "Real-world anomaly detection in surveillance videos," in Proc. IEEE/CVF CVPR, 2018, pp. 6479–6488. DOI: https://doi.org/10.1109/CVPR.2018.00678

[12] R. W. Frischholz and U. Dieckmann, "BioID: a multimodal biometric identification system," Computer, vol. 33, no. 2, pp. 64–68, 2000. DOI: https://doi.org/10.1109/2.820041

[13] S. Yan, Y. Xiong and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in Proc. AAAI Conf. on Artificial Intelligence, vol. 32, no. 1, pp. 7444–7452, 2018. DOI: https://doi.org/10.1609/aaai.v32i1.12328

[14] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz and M. Shah, "Deep learning-based human pose estimation: a survey," ACM Computing Surveys, vol. 56, no. 1, art. 11, pp. 1–37, 2023. DOI: https://doi.org/10.1145/3603618

[15] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in Proc. IEEE/CVF ICCV, 2021, pp. 4975–4986. DOI: https://doi.org/10.1109/ICCV48922.2021.00493

[16] D. Manju, M. Seetha and P. Sammulal, "Early anomalous action detection in surveillance video using MRCNN-LSTM classification," Engineering, Technology & Applied Science Research, vol. 15, no. 4, pp. 25668–25676, 2025. DOI: https://doi.org/10.48084/etasr.10656

[17] Omarov, B., Tursynova, A., & Uzak, M. (2023). Deep learning enhanced internet of medical things to analyze brain computed tomography images of stroke patients. International Journal of Advanced Computer Science and Applications, 14(8).

[18] D. Li, X. Nie, X. Li, Y. Zhang and Y. Yin, "Context-related video anomaly detection via generative adversarial network," Pattern Recognition Letters, vol. 156, pp. 183–189, 2022.

[19] A. Phapale and S. Bhingarkar, "Deep context-aware feature extraction for anomaly detection in CCTV videos," Engineering, Technology & Applied Science Research, vol. 15, no. 2, pp. 21633–21638, 2025. DOI: https://doi.org/10.48084/etasr.9810

[20] W. Ullah, T. Hussain, F. U. M. Ullah, M. Y. Lee and S. W. Baik, "TransCNN: hybrid CNN and transformer mechanism for surveillance anomaly detection," Engineering Applications of Artificial Intelligence, vol. 123, art. 106173, 2023. DOI: https://doi.org/10.1016/j.engappai.2023.106173

[21] R. Nayak, U. C. Pati and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection," Image and Vision Computing, vol. 106, art. 104078, 2021. DOI: https://doi.org/10.1016/j.imavis.2020.104078

[22] L. M. Wastupranata, S. G. Kong and L. Wang, "Deep learning for abnormal human behavior detection in surveillance videos — a survey," Electronics, vol. 13, no. 13, art. 2579, 2024. DOI: https://doi.org/10.3390/electronics13132579

[23] A. Karbalaie, F. Abtahi and M. Sjöström, "Event detection in surveillance videos: a review," Multimedia Tools and Applications, vol. 81, pp. 35463–35501, 2022. DOI: https://doi.org/10.1007/s11042-021-11864-2

[24] L. Shi, Y. Zhang, J. Cheng and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in Proc. IEEE/CVF CVPR, 2019, pp. 12026–12035. DOI: https://doi.org/10.1109/CVPR.2019.01230

[25] T. Gao, J. Chen, H. Wang, S. Zhu, Y. Liu and J. Bai, "Advancing skeleton-based human behavior recognition: multi-stream fusion spatiotemporal graph convolutional networks," Complex & Intelligent Systems, vol. 11, art. 62, 2025. DOI: https://doi.org/10.1007/s40747-024-01743-2

[26] L. Sun, C. Tao, W. Li, Z. Wang, H. Zhao and X. Tang, "Spatial-temporal graph neural ODE networks for skeleton-based action recognition," Scientific Reports, vol. 14, art. 7910, 2024

[27] K. Sun, B. Xiao, D. Liu and J. Wang, "Deep high-resolution representation learning for human pose estimation," in Proc. IEEE/CVF CVPR, 2019, pp. 5693–5703. DOI: https://doi.org/10.1109/CVPR.2019.00584

[28] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using part affinity fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172–186, 2021. DOI: https://doi.org/10.1109/TPAMI.2019.2929257

[29] Ultralytics, "Ultralytics YOLO11," GitHub repository, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[30] T. Liu, X. Hua, S. Qi, H. Liu and L. Teng, "KSL-POSE: a real-time 2D human pose estimation method based on modified YOLOv8-Pose framework," Sensors, vol. 24, no. 19, art. 6249, 2024. DOI: https://doi.org/10.3390/s24196249

[31] Y. Li, P. Xu, X. Liu and J. Chen, "A human pose estimation network based on YOLOv8 framework with efficient multi-scale receptive field and expanded feature pyramid network," Scientific Reports, vol. 15, art. 15081, 2025. DOI: https://doi.org/10.1038/s41598-025-00259-0

[32] Ikram, Z. (2024, May). Hybrid deep neural network for face liveness detection in real-time video. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 188-193). IEEE.

[33] Omarov, B., Baikuvekov, M., Sultan, D., Mukazhanov, N., Suleimenova, M., & Zhekambayeva, M. (2024). Ensemble approach combining deep residual networks and BiGRU with attention mechanism for classification of heart arrhythmias. Computers, Materials, & Continua, 80(1), 341.

[34] A. Shahroudy, J. Liu, T. T. Ng and G. Wang, "NTU RGB+D: a large-scale dataset for 3D human activity analysis," in Proc. IEEE/CVF CVPR, 2016, pp. 1010–1019. DOI: https://doi.org/10.1109/CVPR.2016.115

[35] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan and A. C. Kot, "NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 10, pp. 2684–2701, 2020. DOI: https://doi.org/10.1109/TPAMI.2019.2916873

[36] S. Oh, A. Hoogs, A. Perera et al., "A large-scale benchmark dataset for event recognition in surveillance video," in Proc. IEEE/CVF CVPR, 2011, pp. 3153–3160. DOI: https://doi.org/10.1109/CVPR.2011.5995586

[37] Kulambayev, B., Beissenova, G., Katayev, N., Abduraimova, B., Zhaidakbayeva, L., Sarbassova, A., ... & Shyrakbayev, A. (2022). A Deep Learning-Based Approach for Road Surface Damage Detection. Computers, Materials & Continua, 73(2).

[38] J. Jiao, W. Liu, Y. Mo, J. Jiao, Z. Deng and X. Chen, "Dyn-arcFace: dynamic additive angular margin loss for deep face recognition," Multimedia Tools and Applications, vol. 80, pp. 25741–25756, 2021. DOI: https://doi.org/10.1007/s11042-021-10865-5

[39] A. Gorlov, A. Ulanov, V. Dvorkovich and A. Ershov, "Comparison of information criteria for detection of useful signals in noisy environments," Sensors, vol. 23, no. 4, art. 2133, 2023.

[40] Sultan Mukhamedaly, Kymbat Kabekeyeva, Gulnar Mussabekova, Aliya Kuralbayeva, Bagdat Toibekova, Gulzhan Makashkulova, Batyrkhan Omarov, "A Pedagogical Framework for Ethical Skill Development in Higher Education within Smart Learning Environments", International Journal of Modern Education and Computer Science(IJMECS), Vol.18, No.2, pp. 1-20, 2026. DOI:10.5815/ijmecs.2026.02.01

[41] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. DOI: https://doi.org/10.1023/A:1010933404324

[42] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001. DOI: https://doi.org/10.1214/aos/1013203451

[43] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 785–794. DOI: https://doi.org/10.1145/2939672.2939785

[44] M. Atif, Z. H. Khand, S. Khan, F. Akhtar and A. Rajput, "Storage optimization using adaptive thresholding motion detection," Engineering, Technology & Applied Science Research, vol. 11, no. 2, pp. 6869–6872, 2021. DOI: https://doi.org/10.48084/etasr.3951

[45] Ibrayev, S., Omarov, B., Amanov, B., & Momynkulov, Z. (2024). Development of a deep learning-enhanced lower-limb exoskeleton using electromyography data for post-neurovascular rehabilitation. Engineered Science, 31, 1269.

[46] Raghunath, M. P., Deshmukh, S., Chaudhari, P., Bangare, S. L., Kasat, K., Awasthy, M., ... & Waghulde, R. R. (2025). PCA and PSO based optimized support vector machine for efficient intrusion detection in internet of things. measurement: Sensors, 37, 101806.

[47] Momynkulov, Z., Tursynova, A., Olzhayev, O., Ikramov, A., Ibrayev, S., & Omarov, B. (2025). Three-Dimensional Trajectory Planning for Robotic Manipulators Using Model Predictive Control and Point Cloud Optimization. Computer Modeling in Engineering & Sciences (CMES), 144(4).

[48] A. A. Braga, A. V. Papachristos and D. M. Hureau, "The concentration and stability of gun violence at micro places in Boston, 1980–2008," Journal of Quantitative Criminology, vol. 26, no. 1, pp. 33–53, 2010. DOI: https://doi.org/10.1007/s10940-009-9082-x

[49] D. Weisburd, "The law of crime concentration and the criminology of place," Criminology, vol. 53, no. 2, pp. 133–157, 2015. DOI: https://doi.org/10.1111/1745-9125.12070

[50] S. W. Khan, Q. Hafeez, M. I. Khalid, R. Alroobaea, S. Hussain, J. Iqbal, J. Almotiri and S. S. Ullah, "Anomaly detection in traffic surveillance videos using deep learning," Sensors, vol. 22, no. 17, art. 6563, 2022.

[51] M. Alnowaiser, M. A. Abbas, R. Alroobaea and S. Alzahrani, "An enhanced framework for real-time dense crowd abnormal behavior detection using YOLOv8," Artificial Intelligence Review, vol. 58, art. 182, 2025. DOI: https://doi.org/10.1007/s10462-025-11206-w

[52] D. Jarabo-Amores, R. Gil-Pita, M. Rosa-Zurera and F. Lopez-Ferreras, "Radar detection with the Neyman–Pearson criterion using supervised-learning machines trained with the cross-entropy error," EURASIP Journal on Advances in Signal Processing, vol. 2013, art. 44, 2013. DOI: https://doi.org/10.1186/1687-6180-2013-44

[53] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia and S. Zafeiriou, "ArcFace: additive angular margin loss for deep face recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 5962–5979, 2022. DOI: https://doi.org/10.1109/TPAMI.2021.3087709

[54] S. Chainey, L. Tompson and S. Uhlig, "The utility of hotspot mapping for predicting spatial patterns of crime," Security Journal, vol. 21, no. 1–2, pp. 4–28, 2008. DOI: https://doi.org/10.1057/palgrave.sj.8350066

[55] J. Aldridge and R. Askew, "Delivery dilemmas: How drug cryptomarket users identify and seek to reduce their risk of detection by law enforcement," International Journal of Drug Policy, vol. 41, pp. 101–109, 2017. DOI: https://doi.org/10.1016/j.drugpo.2016.10.010s

[56] M. Al-Masni, M. Al-Dhabyani, A. Alenezi and M. Alshammari, "A real-time intelligent surveillance system for suspicious behavior and facial emotion analysis using YOLOv8 and DeepFace," Engineering Proceedings, vol. 107, no. 1, art. 59, 2025. DOI: https://doi.org/10.3390/engproc2025107059

[57] C. E. Willert, D. M. Mitchell and J. Soria, "An assessment of high-power light-emitting diodes for high frame-rate schlieren imaging," Experiments in Fluids, vol. 53, pp. 413–421, 2012. DOI: https://doi.org/10.1007/s00348-012-1297-1

[58] Le Nguyen, K., Shakouri, M., & Ho, L. S. (2025). Investigating the effectiveness of hybrid gradient boosting models and optimization algorithms for concrete strength prediction. Engineering Applications of Artificial Intelligence, 149, 110568.

[59] Omarov, B. (2025). Deep Learning in Biomedical Image and Signal Processing: A Survey. Computers, Materials, & Continua, 85(2), 2195.

[60] Ikram, Z. (2025). Fourier Transform and Attention Guided Deep Neural Network for Face Anti-Spoofing in Medical Applications. International Journal of Advanced Computer Science & Applications, 16(10).

[61] Ikram, Z. (2025, May). Depth-Guided Neural Network for Robust Face Anti-Spoofing. In 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST) (pp. 1-5). IEEE.