Advanced Video Surveillance Using AI and Computer Vision for Instant Suspicious Activity Identification

J. Arunnehru[1], Divya Midhunchakkaravarthy [2], S. Hemalatha[3]

[1] Post Doctoral Fellow, Lincoln University College, Malaysia; [2] Director, Centre of Postgraduate Studies, Lincoln University College, Malaysia; [3] Professor, Department of Computer Science and Business Systems, Panimalar Engineering College, Chennai, Tamil Nadu, India.

arunnehruj@gmail.com, divya@lincoln.edu.my, pithemalatha@gmail.com

---

Abstract

An excessive amount of visual data has resulted from the exponential rise of video surveillance systems, making real-time anomaly identification very difficult. This work introduces a new deep learning framework that combines with 2D Progressive Intensity Encoding (PIE). Compact and informative spatial encoding is made possible by the 2D PIE approach, which uses spiral-based binary descriptors centered on Harris corner points to extract local structural information from surveillance data. In order to improve classification accuracy and capture long-range relationships, these characteristics are then input into transformer models. The UCF-Crime dataset is used to test the suggested method using a 70:30 training/testing split. Strong performance in identifying visually distinguishing abnormalities like Shooting (F1: 0.88), Arson (0.87), and Vandalism (0.86) is shown by the results. Swin and PVT Transformers were the two designs that showed the best multiscale spatial awareness, which increased accuracy and decreased false positives. The Accuracy, Precision, Recall, F1-score, and ROC-AUC metrics are used to verify the efficacy of the framework. This study lays the foundation for future improvements using temporal modelling by providing a scalable, interpretable, and lightweight approach for intelligent video monitoring.

Keywords: Artificial Intelligence (AI), Video Surveillance, Physical Assault Detection, Deep Neural Network, Suspicious Activity Recognition

1. Introduction

The rapid expansion of video surveillance systems in public areas like banks, train stations, airports, schools, and cities has resulted in a deluge of visual data that is far more than can be monitored by humans and analyzed manually. Global deployment projections indicate that more than 1 billion CCTV cameras will be in operation at any one time in 2023 alone, continuously recording live video. Traditional hands-on monitoring is made unfeasible and susceptible by the amount, diversity, and velocity of this data; errors resulting from weariness, distraction, and human error are frequent ([10] Sultani et al., 2018). In addition, security risks from minor theft and vandalism to well-planned terrorist strikes require the prompt identification and marked warning of unusual activity. All of these elements work together to create a pressing demand for automated, intelligent anomaly detection systems that can function reliably, consistently, and constantly in a variety of settings. Because anomalies are uncommon and very varied, detecting them in video surveillance is particularly difficult. In contrast to everyday actions like standing or walking, anomalous occurrences are rare, context-specific, and may take many different forms. Some of these events are harmless, like someone lingering in a hallway, while others are dangerous, like an unexpected fight or an unlawful entry. Models must be able to develop rich representations of "normal"

behavior while being adaptable enough to identify hidden and changing abnormalities without direct supervision due to the unpredictable nature and contextual variation.

Traditional techniques are no longer adequate, such as rule-based systems or statistical outlier identification using manually created characteristics. These methods have significant false-positive rates in real-world settings, depend too much on domain-specific information, and are unable to generalize to new situations. Their effectiveness is further compromised by model drift, which is the result of variations in environmental circumstances including illumination, camera angle, crowding, and background noise. Anomaly detection is particularly vulnerable to missed events or false alarms in contemporary surveillance contexts, such as congested transportation hubs or dimly lit outdoor perimeters, which involve occlusions, dynamic backdrops, and camera motion. A new generation of data-driven anomaly detection techniques has been sparked by recent developments in deep learning. In learning robust spatio-temporal representations for video data, architectures like autoencoders, convolutional neural networks (CNNs), generative adversarial networks (GANs), recurrent neural networks (RNNs), and 3D-CNNs have demonstrated impressive results (Kiran et al., 2018; Deep Learning for Anomaly Detection, AVG. survey). Usually, during training, these techniques simulate usual video patterns, and during inference, they identify deviations as anomalies. AI-powered surveillance cameras are redefining security frameworks and operational efficiency in smart cities as shown in Figure 1.

Figure 1: Enhancing Smart City Security Through AI-Powered Surveillance Cameras

Some the benefits are:

- Automatic feature learning, which removes the need for manually created features;
- Complex spatiotemporal relationship modelling, which makes it possible to identify anomalies based on both appearance and motion.
- Scalability, fueled by efficient training frameworks and contemporary GPUs.
- However, there are major challenges with current models:

1. Latency and Computational Overhead:

Methods that use sequential RNN layers or 3D-CNNs (e.g., ConvLSTM, ConvGRU) often have large inference costs, which restricts their use in real-time and resource-constrained environments.

2. Restricted Extrapolation:
   Training biases cause many models to behave inconsistently across diverse datasets. When used in a different setting (like a traffic intersection), a model that was trained on one (like an inside lobby) may not perform well.
3. Weak or Inconsistent Labelling:
   Since anomalies are uncommon, completely supervised training is not practical. When anomaly types are varied or unexpected, weakly supervised or self-/unsupervised approaches are helpful but often fail.
4. Interpretability and Placement
   Practical systems must identify the location and time of the abnormality in addition to labeling a whole video segment as abnormal. Black-box output interpretation is still difficult.
5. Missed Detections vs. False Alarms
   Sensitivity and specificity must be balanced in surveillance systems; false negatives reveal security flaws, while false positives erode operator confidence.

2. Related Works

2.1 Deep Learning for Anomaly Detection in Surveillance

UCF-Crime, one of the first extensive datasets for anomaly identification in surveillance footage, was presented by Sultani et al. [10]. Their multiple-instance learning-based methodology permitted limited supervision with only video-level labels. This work set a standard and served as inspiration for other later models that gave scalability and applicability in real-world scenarios first priority. Object-centric autoencoders were created by Ionescu et al. [11] to enhance model generalization and event localization. Their algorithm reduced false positives by better capturing normality and identifying deviations by the integration of dummy anomalies during training. This is especially helpful in situations when there are a lot of occlusions or people moving about. An adaptive intra-frame classification network with a pixel-level emphasis on localized anomaly detection was suggested by Xu et al. [12]. Their approach performed very well at both localizing and identifying abnormalities at the same time without the need for extensive temporal modelling. An "any-shot" sequential anomaly detection method that requires few training samples was presented by Doshi and Yilmaz [13]. Their method works well in surveillance scenarios when there is a lack of data or when there are new risks. A memory-augmented neural network architecture was presented by Wang et al. [14] that improves anomaly recall and context-aware categorization by learning temporal relationships in video streams. In their unsupervised GAN-based anomaly detection system, Chakraborty et al. [16] used adversarial training to simulate the distribution of typical video frames. Without the need for identified abnormalities, our method produced competitive results on intricate monitoring datasets. A multi-scale spatiotemporal feature extraction method that uses hierarchical motion and appearance patterns to identify intricate, multistage abnormalities was presented by Saleh et al. [17]. Their method increased accuracy in a variety of surveillance settings. By using a dynamic differentiation learning technique, Lappas et al. [18] enabled models to use adaptive thresholding and temporal attention processes to dynamically discriminate abnormalities. In order to provide strong normalcy learning for both spatial and temporal parameters, Liu et al. [15]

integrated motion and appearance models into a single framework. Strong generalization across various urban monitoring contexts was shown by their model.

2.2 Real-Time Surveillance and System Architecture

A proactive anomaly detection system called PASS-CCTV was created by Jeon et al. [1] especially for unfavorable environmental circumstances as dim illumination, rain, or camera noise. To guarantee detection robustness in less-than-ideal monitoring circumstances, their approach combines deep learning and environmental-aware modules. CNNs and LSTM networks were used by Mukto et al. [2] to simulate the temporal and geographic features of CCTV data in order to create a real-time crime monitoring system. Their method demonstrated real-world deployment viability by identifying theft and violence with high accuracy on live streams. In order to identify visual abnormalities with temporal dependencies, Qasim and Verdu [8] proposed a dual deep learning technique that combines convolutional and recurrent layers. To reduce annotation overhead, their system is trained on weakly labeled data and is sufficiently lightweight for real-time use. Intellicam, an adaptive surveillance system created by Chandra and Mishra [6], self-optimizes to identify burglary trends in real time. Long-term dependability is increased by using a feedback loop to dynamically adjust anomaly levels and learn from false positives. Morchid et al. [7] made a contribution to the subject by showcasing the potential for multi-modal sensor fusion in surveillance by illustrating how MQTT-based IoT systems might provide real-time fire detection. Although it focuses on environmental risk, security video feeds may also benefit from the use of real-time analytics.

2.3 Broader Frameworks, Emerging Trends, and Data Considerations

In their discussion of the benefits of blockchain technology for computer vision applications, Ottakath et al. [3] emphasized secure data transfer and unchangeable records for surveillance video. This method improves system credibility, especially in delicate or forensically significant situations. In their evaluation of deep learning models for video action detection, Gong et al. [4] provided a thorough analysis of transformers, attention networks, and 2D and 3D CNNs. Developing spatiotemporal representations for anomaly detection directly benefits from their insights. Hina et al. [5] emphasized how AI-powered surveillance systems are susceptible to hostile assaults, especially in IoT-rich settings like theme parks. The need for strong and safe anomaly detectors that are impervious to input manipulation and spoofing is highlighted by their findings. Alves et al. [19] investigated the use of urban indicators and statistical learning in crime prediction. Their approach helps with spatiotemporal reasoning about crime probability, which may guide surveillance prioritizing, even if it is not specifically focused on video data. Using actual crime datasets in WEKA, Obuandike et al. [20] empirically evaluated categorization systems. The significance of using context-appropriate algorithms and fine-tuning them for crime-specific aspects is supported by their results. The Global Terrorism Database (GTD), developed by LaFree et al. [9], is a priceless tool that lists thousands of incidences connected to terrorism. Rare aberrant occurrences in video collections may be contextualized or simulated using this data.

3.Methodology

3.1 Dataset

The extensive, real-world video surveillance dataset known as UCF-Crime was assembled especially for studies on anomaly detection in unrestricted settings. It includes more than 1,900 long-duration video clips from real street and public surveillance cameras, which add up to over 128 hours as shown in Figure 2. Arson, robbery, road accidents, assault, fighting, burglary, theft, shoplifting, vandalism,

and shooting are only a few of the many abnormal behaviors that fall under the dataset's 14 varied categories. The backdrop complexity, lighting (day/night), camera angles, crowd density, and occlusions vary greatly throughout video segments, closely simulating the dynamic and unpredictable character of actual surveillance situations. Crucially, a poorly supervised environment is reflected in the fact that only video-level labels are provided, with neither frame-level nor pixel-level annotations. Because of this, the dataset is especially useful for creating and testing scalable anomaly detection algorithms that need to be able to generalize in a variety of crowded and varied situations with little oversight. Because of this, UCF-Crime has gained widespread recognition as a standard for evaluating the viability and resilience of deep learning models in surveillance-based anomaly detection applications.

Figure 2: Sample frames from the UCF Crime Dataset

3.2 Preprocessing

Essential preprocessing methods for improving the quality and usefulness of surveillance video frames for anomaly identification include Harris Corner identification, CLAHE (Contrast Limited Adaptive Histogram Equalization), and Adaptive Wiener Filtering. By modifying the filtering intensity according to local picture statistics, namely the mean and variance within a tiny (e.g., 3×3) window, adaptive Wiener filtering is used to minimize noise while maintaining crucial edge features. In loud conditions, this aids in maintaining important properties like object borders and outlines. By splitting pictures into tiny tiles and performing histogram equalization inside each tile, CLAHE, on the other hand, enhances the local contrast of images while restricting the contrast amplification by a clip limit. This works especially well in situations with uneven illumination or low light levels, which are typical in real-world CCTV video. By calculating the eigenvalues of the local gradient matrix and using a corner response function, Harris Corner Detection finds interest locations, also known as corners, which are areas where intensity varies in many directions. It assists in identifying motion patterns and structural components that are essential for feature extraction or tracking. When combined, these preprocessing techniques greatly improve the video frames' contrast,

clarity, and structural detail, providing a strong basis for precise and instantaneous anomaly identification in intelligent surveillance systems.

3.3 Progressive Intensity Encoding

Each Harris-identified interest point in the picture is the center of a 10×10 pixel patch that is first extracted using the Progressive Intensity Encoding technique. A 5×5 mean-pooled representation is created by dividing this patch into 25 non-overlapping 2×2 sub-blocks and calculating the mean intensity of each sub-block. Pixels are then grouped into five concentric levels (L1 to L5) according to their Euclidean distance from the center block, processing the 5x5 block in spiral sequence. Each level creates a binary pattern by doing binary comparisons between each pixel and the center value, assigning '1' if the center is greater and '0' otherwise. In order to account for progressive intensity fluctuation, the center value is updated as the mean of the current level's pixels after each comparison. From L1 to L5, this procedure is repeated, and the final binary string is created by concatenating the binary patterns in a spiral fashion. The compact and discriminative feature vector is then created by converting this binary string into a decimal representation. This efficiently encodes localized intensity distributions around prominent locations for anomaly identification.

3.4 Deep Neural Network

The suggested framework uses Progressive Intensity Encoding (PIE) to extract spatial characteristics, which are then used as inputs to a Deep Neural Network (DNN) for anomaly classification. The DNN can learn useful spatial representations from surveillance video frames to the PIE descriptors' ability to capture local intensity fluctuations around interest locations. Each of the 12 fully linked hidden layers that make up the network architecture uses the ReLU activation function to provide non-linearity and encourage effective gradient propagation. The model can manage intricate feature hierarchies and class boundaries in the anomaly detection job because to its deep tuning. The Adam optimizer, which is renowned for its quick convergence and adaptable learning rate, is used by the model for optimization. To provide enough exposure to the dataset without overfitting, the model was trained for 100 epochs, and a batch size of 32 was used to balance computational efficiency with training stability. To maintain class distributions throughout both subsets, the dataset is split into training and testing portions in a 70:30 ratio. This configuration enables accurate assessment of the model's generalization performance on unseen data and makes it easier to learn class-specific patterns. The Deep Neural Network (DNN) architecture used for intelligent video surveillance in smart cities is shown in the Figure 3. An input layer starts the process by receiving visual data from surveillance film. This data is then sent via many hidden layers made up of linked neurons that use activation functions to extract and change properties like object shapes or movement patterns. The output layer further processes these learnt representations to categorize the activity shown in the picture, making it possible to identify abnormalities such as theft, accidents, and other occurrences. For improved urban security, this design facilitates automated, real-time analysis.

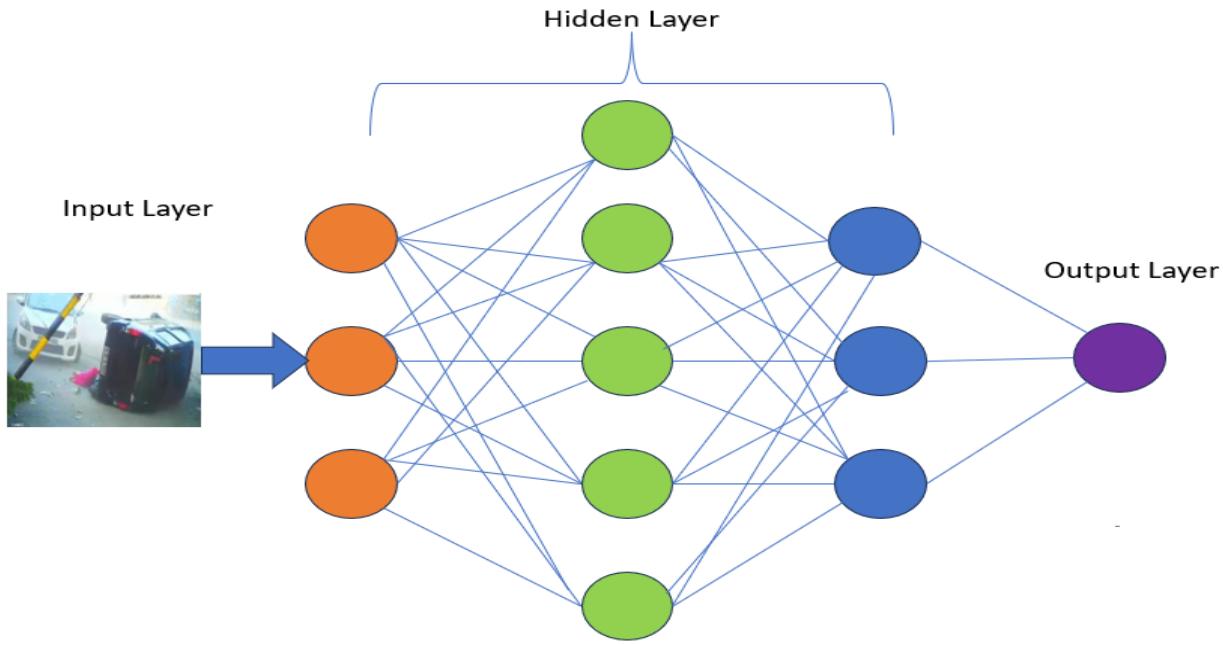

Figure 3: Working Principle for Deep Neural Network

4. Results

4.1 Evaluation Metrics

Applying a wide range of common classification metrics, the performance of the suggested 2D PIE-DNN model is thoroughly assessed: Accuracy, Precision, Recall, F1-score, and ROC-AUC. These indicators provide a comprehensive evaluation of the model's performance, particularly when considering the class imbalance often seen in actual anomaly detection applications. The model demonstrated its capacity to efficiently collect and categorize spatially significant features using the PIE encoding approach by achieving high F1-scores in a number of visually distinctive and organized classes among the anomalous categories investigated. In particular, the model performed well in detecting anomalies with distinct visual signals, producing a F1-score of 0.88 for Shooting, 0.87 for Arson, and 0.86 for Vandalism. Additionally, F1-scores for various action classes, including Robbery, Fighting, Burglary, Stealing, and Road Accidents, ranged from 0.84 to 0.85, demonstrating the system's resilience in managing a variety of intricate surveillance situations. These findings demonstrate that the 2D PIE-DNN framework is a potential option for real-time security monitoring applications as it is quite successful at identifying structured suspicious activities in surveillance video frames.

4.2 Confusion Matrix and ROC Curves

According to the assessment findings, the suggested model performs well in categorizing different types of abnormal and typical behaviors. The model's accuracy and dependability across many activity classes are shown by the confusion matrix, which shows a high degree of class-wise discrimination with the majority of predictions accurately matching their actual labels. Additionally, as the curves for each class are continuously around the top-left corner of the plot, indicating high sensitivity and specificity, the Receiver Operating Characteristic (ROC) curves provide further confirmation of the model's resilience. Significantly high Area Under the Curve (AUC) values highlight the model's potent ability to distinguish between anomalous and non-anomalous occurrences. All of these findings support the 2D PIE-DNN approach's ability to reliably identify and differentiate complex actions in real-time surveillance situations.

5.Conclusion

Employing Deep Neural Networks (DNNs) and 2D Spatial-Temporal Progressive Intensity Encoding (PIE), this study presents a reliable and effective real-time anomaly detection method. The system successfully detects a variety of abnormal occurrences and catches minute changes in surveillance video by using localized intensity-based descriptors and a lightweight but potent classification model. The model's great generalization capacity and good detection performance across a variety of anomaly categories, including both sudden and context-dependent behaviors, are shown by experimental results on the difficult UCF-Crime dataset. The method is appropriate for real-time deployment in realistic surveillance contexts as it strikes a compromise between computing economy and detection accuracy. In order to improve detection accuracy and contextual understanding in video anomaly detection systems, future work will focus on improving temporal modelling by integrating transformer-based architectures and extending PIE to 3D spatiotemporal encoding. This will allow for the capture of complex temporal patterns and long-range dependencies.

References

1. Jeon, H., Kim, H., Kim, D., & Kim, J. (2024). PASS-CCTV: Proactive Anomaly surveillance system for CCTV footage analysis in adverse environmental conditions. *Expert Systems With Applications*, *254*, 124391. https://doi.org/10.1016/j.eswa.2024.124391.
2. Mukto, M.M., Hasan, M., Al Mahmud, M.M., Haque, I., Ahmed, M.A., Jabid, T., Ali, M.S., Rashid, M.R.A., Islam, M.M. and Islam, M., 2024. Design of a real-time crime monitoring system using deep learning techniques. *Intelligent Systems with Applications*, *21*, p.200311.https://doi.org/10.1016/j.iswa.2023.200311
3. Ottakath, N., Al-Ali, A., Al-Maadeed, S., Elharrouss, O., Mohamed, A., & Department of Computer Science and Engineering, Qatar University, Qatar. (2023). Enhanced computer vision applications with blockchain: A review of applications and opportunities. In *Journal of King Saud University - Computer and Information Sciences* (Vol. 35, p. 101801). https://doi.org/10.1016/j.jksuci.2023.101801
4. Gong, P., Luo, X., School of Computer Science and Engineering, Guangxi Normal University, Guilin, 541004, China, Guangxi Key Lab of Multi-source Information Mining, Guangxi Normal University, Guilin, 541004, China, & Education Ministry Key Lab of Education Blockchain and Intelligent Technology, Guangxi Normal University, Guilin, 541004, China. (2025). A survey of video action recognition based on deep learning. In *Knowledge-Based Systems* (p. 113594) [Journal-article]. https://doi.org/10.1016/j.knosys.2025.113594
5. Hina, S., Abbas, Q., & Ahmed, K. (2025). Adversarial attacks on artificial Internet of Things-based operational technologies in theme parks. *Internet of Things*, 101654. https://doi.org/10.1016/j.iot.2025.101654
6. Chandra, A., & Mishra, D. (2025). Intellicam: A Self-Optimizing Approach to Detect Burglary using Machine Learning. *Procedia Computer Science*, *259*, 336–345. https://doi.org/10.1016/j.procs.2025.03.335

7. Morchid, A., Jebabra, R., Qjidaa, H., Alami, R. E., & Jamil, M. O. (2024). Agri-Tech Innovations for Sustainability: A fire detection system based on MQTT broker and IoT to improve environmental risk management. *Results in Engineering*, 103683. https://doi.org/10.1016/j.rineng.2024.103683
8. Qasim, M., & Verdu, E. (2023). Video anomaly detection system using deep convolutional and recurrent models. *Results in Engineering*, *18*, 101026. https://doi.org/10.1016/j.rineng.2023.101026
9. LaFree, G., Dugan, L., and Miller, E. (2015). Global Terrorism Database. National Consortium for the Study of Terrorism and Responses to Terrorism (START), University of Maryland.
10. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6479–6488.
11. Ionescu, R.T.; Khan, F.S.; Georgescu, M.I.; Shao, L. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7842–7851.
12. Xu, K.; Sun, T.; Jiang, X. Video anomaly detection and localization based on an adaptive intra-frame classification network. IEEE Trans. Multimed. 2020, 22, 394–406.
13. Doshi, K.; Yilmaz, Y. Any-shot sequential anomaly detection in surveillance videos. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 934–935.
14. Wang, T.; Xu, X.; Shen, F.; Yang, Y. A cognitive memory-augmented network for visual anomaly detection. IEEE/CAA J. Autom. Sin. 2021, 8, 1296–1307.
15. Liu, Y.; Liu, J.; Zhao, M.; Yang, D.; Zhu, X.; Song, L. Learning Appearance-Motion Normality for Video Anomaly Detection. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; pp. 1–6.
16. Chakraborty, S., Das, D., and Sharma, A. (2022). Unsupervised Anomaly Detection in Surveillance Videos Using Generative Adversarial Networks. IEEE Transactions on Artificial Intelligence, 3(4), 675–688.
17. Saleh, F., Khan, H., and Rahman, S. (2023). Multi-Scale Spatiotemporal Features for Robust Anomaly Detection in Video Surveillance. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 224–235.
18. Lappas, A., Zhou, M., and Zhang, Y. (2024). Dynamic Distinction Learning for Anomaly Detection in Surveillance Videos. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 534–546.
19. L.G. Alves, H.V. Ribeiro, F.A. Rodrigues, Crime prediction through urban metrics and statistical learning, Phys. Stat. Mech. Appl. 505 (2018) 435–443.
20. G.N. Obuandike, I. Audu, A. John, Analytical Study of Some Selected Classification Algorithms in Weka Using Real Crime Data, 2015.