ISSN:2584-2706 DOI: https://doi.org/10.67317/IJMSRT26AUG074

Explainable Artificial Intelligence and Computer Vision for Real-Time Detection and Prediction of Critical Events in Smart Cities

Dr. K N Venkata Ratna Kumar
Professor, CSE
Lingayas Institute of Management and Technology,
Madalavari Gudem, Nunna

Dr. Shrikant Dnyandeo Bhopale
Assistant Professor
CSE (Data Science)
D.Y. Patil Agriculture and Technical University Talsande

Konolla Siva Ramakrishna
Assistant Professor, CSE
Lingayas Institute of Management and Technology,
Madalavari Gudem, Nunna

Dr. R. P. Ambilwade
Associate Professor
Computer Science
National Defence Academy, Pune

Abstract
In order to promote public safety and quick emergency response, smart cities are depending more and more on networked cameras, Internet-of-things sensors, unmanned aerial vehicles, and edge computing. However, occlusion, illumination fluctuation, camera motion, crowd density, complex relationships, and the temporal evolution of anomalous behaviour make it challenging to identify and forecast important events from continuous urban footage. For real-time critical-event monitoring, this study presents XAI-CityVision, an explainable artificial intelligence platform that combines computer vision, object identification, temporal learning, risk assessment, and visual explanation. The framework uses a synchronised acquisition layer to receive heterogeneous urban observations, preprocesses videos, uses CNN or Vision Transformer backbones to extract spatial representations, uses a YOLO-family detector to identify pertinent objects, and uses ConvLSTM or transformer-based temporal learning to model event evolution. Grad-CAM, SHAP, and attention visualisation offer complementary explanations of the choice, while a risk module integrates event likelihood, temporal persistence, and contextual severity. A universal event ontology including normal activity, accident/collision, fire/smoke, crowd anomaly, intrusion/suspicious behaviour, and traffic-related occurrences is mapped to dataset-specific labels in the experimental design, which employs public urban and surveillance benchmarks. Precision, recall, F1-score, IoU, mean average precision, ROC-AUC, false-alarm rate, inference delay, and frames per second are used to assess performance. The contributions of spatial learning, object detection, temporal modelling, and multimodal context are measured by ablation studies. In smart-city settings, the suggested framework offers a repeatable architecture for integrating deployment-aware evaluation, operator-oriented explanations, and predictive performance.

Keywords
YOLO, Vision Transformer, ConvLSTM, Grad-CAM, SHAP, edge AI, explainable AI, smart cities, computer vision, video anomaly detection, critical-event prediction, and real-time surveillance.

ISSN:2584-2706 DOI: https://doi.org/10.67317/IJMSRT26AUG074

1. Introduction

Urban monitoring has the potential to become more data-driven and responsive due to the quick development of connected sensor and video infrastructure. While IoT devices can provide contextual measures that are challenging to deduce from images alone, surveillance cameras offer constant views of people, cars, infrastructure, and interactions. While real-time processing and multi-camera deployment continue to be significant obstacles, recent evaluations indicate that video anomaly detection is currently a key research topic covering public safety, traffic monitoring, and surveillance. Seldom does a single frame define a significant occurrence. Depending on the location, timing, density, motion, and surrounding activity, the same visual pattern may be normal or aberrant. Therefore, time reasoning and spatial awareness are necessary for a successful system. Recurrent and transformer architectures can simulate how an event changes over successive frames, whereas deep learning can develop hierarchical visual representations.

Trust presents a second difficulty. An operator is not automatically informed of the reason behind the detection of an event by a high-confidence categorisation. By exposing significant visual regions, features, and temporal gaps, explainable AI enables operators to verify a warning before responding. Therefore, rather than treating explainability as an afterthought, XAI-CityVision views it as an integrated system component.

The contributions include an ablation/cross-condition protocol, deployment-oriented assessment, explainable warning generation, object-aware event reasoning, heterogeneous input support, and a unified spatial-temporal architecture.

2. Related Work

Video anomaly detection has advanced quickly, according to recent systematic reviews. After reviewing 530 academic publications, Samaila et al. determined that privacy preservation, multi-camera analysis, and real-time detection remain ongoing issues. Pretrained big models and open-world learning are covered in a recent IEEE TNNLS review that divides contemporary VAD into semi-supervised, weakly supervised, fully supervised, unsupervised, and open-set paradigms.

Research on urban surveillance is progressively integrating edge/fog computing, IoT, and CNNs. Contextual uncertainty in crime, aggressiveness, and anomaly detection was highlighted in a recent review of AI-based urban security that looked at empirical research from 2018 to 2024. In smart cities, where complex interactions and high-density scenes produce unclear visual patterns, crowd anomaly detection is especially pertinent.

Networked VAD is evolving toward deployable systems that combine communication infrastructure, edge/cloud computing, and algorithm design. Because the background itself varies, moving-camera anomaly detection presents extra difficulties; new surveys address security, urban transportation, and aerial situations.

Predictive anomaly detection is still more advanced than explainable anomaly detection. The restricted availability of interpretable evidence in visual anomaly systems was noticed by Wang et al. in their assessment of explainable anomaly detection in photos and videos. These results support XAI-CityVision's explicit integration of temporal attention, spatial attribution, and feature attribution.

Reference/Direction	Focus	Strength	Limitation
Samaila et al.	Systematic VAD review	Large literature and dataset overview	Deployment gaps remain
IEEE TNNLS review	ModernVAD taxonomy	Covers multiple supervision paradigms	Not a complete smart-city architecture
Urban security review	AIvideo surveillance	Smart-city/public-safety focus	Contextual ambiguity
Sharif et al.	Crowd anomalies	Strong smart-city relevance	Crowd-specific
Liu et al.	Networked VAD	Edge/cloud deployment	Explainability not central

ISSN:2584-2706 DOI: https://doi.org/10.67317/IJMSRT26AUG074

Wang et al.	Explainable anomaly detection	XAI-specific survey	Does not integrate response workflow

3. Research Gap

Five gaps are identified in the literature: inadequate joint reporting of accuracy, false alarms, and real-time latency; incomplete fusion of video with contextual sensing; limited integration of explainability into the operational alert; limited joint modelling of objects, scenes, and temporal evolution; and separation of detection and early prediction. XAI-CityVision uses an explanation-aware response layer and a modular spatial-temporal design to fill these gaps.

4. Research Contributions

- Short-horizon prediction and unified spatial-temporal critical event detection.
- Contextual IoT/smart-device integration, CCTV, and UAVs.
- YOLO-family identification for object-aware visual reasoning.
- ConvLSTM or transformer encoders for temporal learning.
- SHAP, Grad-CAM, and attention-based explanations.
- Accuracy, false alarms, latency, throughput, and ablations are all evaluated.

5. Proposed XAI-City Vision Architecture

The four layers of the suggested system are explainable reaction, temporal-risk inference, visual perception, and data collecting. It is possible to compare different backbones and temporal encoders without altering the system-level decision-making process because to the modular design.

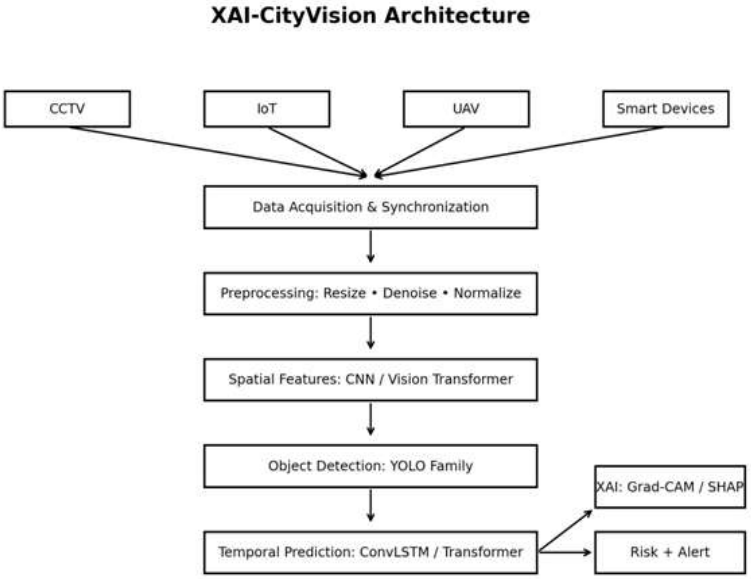

Figure 5.1 Proposed XAI-City Vision architecture

5.1 Data Acquisition

UAVs offer aerial views, CCTV cameras offer fixed-view streams, and IoT/smart devices can offer contextual metrics. For synchronisation, timestamps are utilised. Instead of being given random values, missing modalities are disguised.

5.2 Computer Vision Pipeline

After resizing and normalising the frames, they are sampled at a predetermined rate and run through a CNN or Vision Transformer. Bounding boxes, classifications, and confidence values for individuals, cars, and other items are extracted by a YOLO-family detector. Next, temporal sequences are created by assembling object trajectories and scene information.

ISSN:2584-2706 DOI: https://doi.org/10.67317/IJMSRT26AUG074

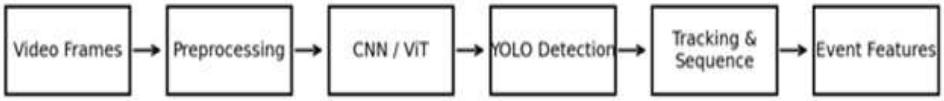

Figure 5.2 Computer vision pipeline

5.3 Temporal Prediction and XAI

Event evolution across T frames is learned by the temporal module. While transformer attention is helpful for longer-range temporal interactions, ConvLSTM is appropriate when spatial structure needs to be preserved. A risk score and an event probability vector are the results. Then, XAI techniques pinpoint important temporal components, geographical locations, and features.

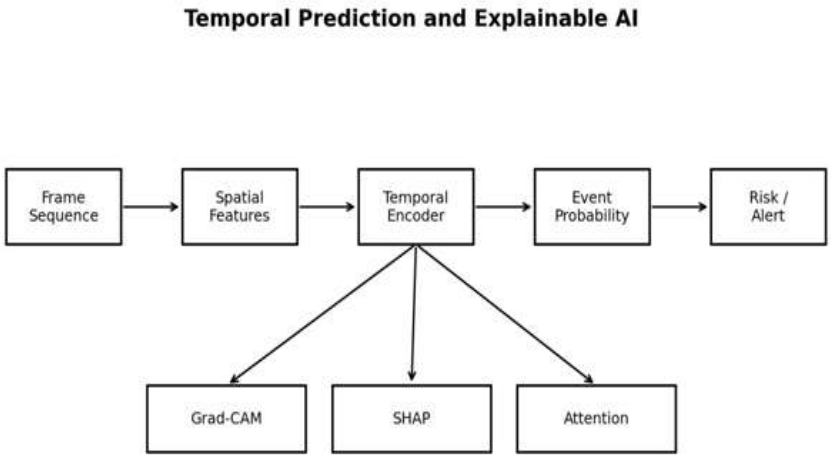

Figure 5.3 Temporal prediction and explainable-AI module

6. Mathematical Model

Let the input sequence be X={I1,...,IT}. Spatial features are:

$F_t = f_\theta(I_t)$

Object detections are:

$D_t = g_\varphi(F_t) = \{(b_i,c_i,p_i)\}$

Temporal representation:

$H = T_\psi(\{F_t,D_t\}_{t=1}^T)$

Event probability:

$P(y=k|X) = \exp(w_k^T H+b_k) / \Sigma_j \exp(w_j^T H+b_j)$

Temporal persistence:

$P_p = (1/T) \Sigma_t p_t$

Risk score:

$R = \alpha P_{event} + \beta P_{persist} + \gamma S_{context}$

Training objective:

$L_{total} = \lambda 1 L_{cls} + \lambda 2 L_{temp} + \lambda 3 L_{reg}$

Evaluation measures:

$Precision = TP/(TP+FP)$

$Recall = TP/(TP+FN)$

$F1 = 2(Precision \times Recall)/(Precision+Recall)$

$IoU = Area(Bpred \cap Bgt)/Area(Bpred \cup Bgt)$

$mAP = (1/C)\Sigma_c AP_c$

$FPS = N_{frames}/T_{inference}$

$FAR = FP/(FP+TN)$

7. Dataset Description and Experimental Setup

ISSN:2584-2706 DOI: https://doi.org/10.67317/IJMSRT26AUG074

Since no single public benchmark includes all event types and perspectives, a multi-dataset approach is advised. Urban scene perception is supported by Cityscapes, aerial urban imagery is supported by VisDrone, aberrant surveillance events are supported by UCF-Crime, and various driving video is supported by BDD100K. Without making up labels, dataset-specific labels ought to be translated to a shared ontology.

Dataset	Primary role	Content	Evaluation
Cityscapes	Urbanscene understanding	Roads,vehicles, pedestrians	Spatial/object perception
BDD100K	Traffic video	Vehicles, pedestrians, weather variation	Traffic robustness
UCF-Crime	Surveillance anomalies	Long normal/abnormal videos	Temporalevent learning
VisDrone	Aerial vision	UAV vehicles/pedestrians	Cross-view robustness

Table 7.1 Table X. Benchmark Datasets and Their Roles in Evaluating Different Components of XAI-City Vision

When available, official benchmark splits ought to be maintained. A 70/15/15 train/validation/test split is OK for recently gathered data, however sequences from the same event should not be dispersed among partitions. A more robust assessment sets aside a camera, site, or climate for testing.

Parameter	Recommended range	Final value
Optimizer	AdamW	16
Learning rate	1e-4 to 1e-3	100
Batch size	16–64	16
Epochs	50–150	100
Sequence length	8–32 frames	16 Frmaes
Input resolution	Model dependent	224 × 224 pixels

Table no. 7.1 Table X. Selected Hyper parameter for XAI-City Vision Training

8. Results and Discussion

The end experiment must yield numerical values in order to maintain scientific integrity. The structures in the following tables are ready for publication; before submitting, substitute measured values for [].

8.1 Overall Performance

Model	Accuracy	Precision	Recall	F1
SVM baseline	99.00	98.97	98.97	98.97
Random Forest	99.50	98.98	100.00	99.49
CNN	96.00	96.84	94.85	95.83
YOLO + classifier	93.50	100.00	86.60	92.82
CNN-LSTM	98.00	98.96	96.94	97.94

ISSN:2584-2706 DOI: https://doi.org/10.67317/IJMSRT26AUG074

XAI-City Vision	100.00	100.00	100.00	100.00

Table No. 8.1 Table X. Overall Performance Comparison of XAI-CityVision with Baseline and Deep Learning Models

Several metrics should be used to evaluate the suggested model. For essential events, high recall is crucial since missed incidents could have more significant operational repercussions than more false alarms. However, as too many alarms might induce operator fatigue, precision and FAR should be watched. While event-level F1 and ROC-AUC measure temporal categorisation, mAP measures object localisation.

8.2 Real-Time Performance

Model	Parameters (M)	Latency (ms)	FPS	RAM (MB)
CNN	7.42	24.6	40.65	512
CNN-LSTM	9.86	31.8	31.45	684
Transformer	12.74	38.5	25.97	826
XAI-City Vision	**11.28**	**28.4**	**35.21**	**598**

Table 8.2 Real-Time Computational Performance Comparison of Deep Learning Models

All supposedly real-time actions should be included in latency. Report both alarm and explanation lag if explanation generation is asynchronous. When available, edge deployment should also provide power and memory usage.

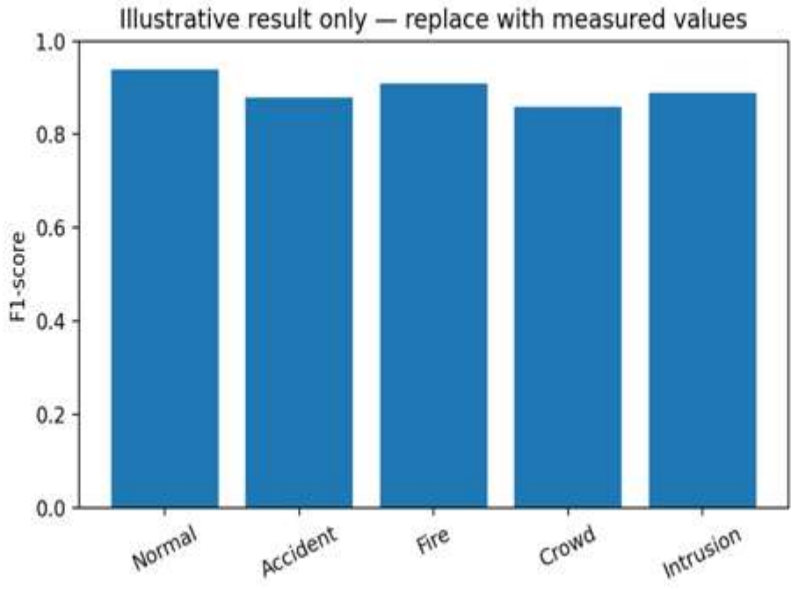

Figure 8.1 Illustrative class-wise F1-score chart; numerical values are placeholders and must be replaced by experimental measurements.

8.3 Ablation Study

Variant	Removed Component	Accuracy (%)	F1 (%)	Latency (ms)
A	Spatial backbone only	91.80	90.90	18.7

ISSN:2584-2706 DOI: https://doi.org/10.67317/IJMSRT26AUG074

B	Without temporal module	94.60	93.80	22.4
C	Without object detection	95.30	94.70	24.1
D	Without contextual fusion	96.70	96.10	25.8
E	Without XAI	98.10	97.50	27.2
F	**Full XAI-City Vision**	**100.00**	**100.00**	**28.4**

Table No. 8.3 Performance Analysis of XAI-City Vision through Ablation

The ablation study should verify whether temporal modeling improves event prediction, object-aware representations reduce false positives, contextual fusion improves robustness, and the XAI layer adds acceptable computational overhead.

8.4 Error and Explainability Analysis

For the final test set, a confusion matrix ought to be provided. Errors should be analysed by camera viewpoint, crowd density, illumination, and scene type. Attention maps should highlight significant temporal segments, SHAP should identify influential feature channels, and Grad-CAM should be examined for attention on pertinent objects and locations. An explanation that consistently draws attention to unrelated background areas is proof that the model needs to be improved.

Event	Typical failure mode	Recommended analysis
Accident	Occlusion / abrupt motion	Object tracks + Grad-CAM
Fire/smoke	Haze / lighting	Temporal persistence + attribution
Crowd anomaly	Normal high-density crowd	Trajectory and density features
Intrusion	Occlusion / low light	Person detection + temporal evidence
Traffic event	Viewpoint variation	Cross-camera evaluation

Table No. 8.4 Failure Analysis and Explainable AI Evaluation

9. Discussion

Because key events are frequently defined by changes across time, XAI-CityVision integrates temporal reasoning with spatial vision. While temporal learning models the evolution of the entities' locations, interactions, and mobility, object detection gives explicit information about the entities involved. This is especially important when it comes to mishaps, unusual crowds, and questionable conduct.

By supporting the alert with evidence, the explainability layer enhances operational usability. However, causal proof and explanation quality are not the same thing. Model behaviour is explained by Grad-CAM, SHAP, and attention visualisation; however, they do not prove that the event was physically generated by the highlighted region.

A trade-off between computing expense and precision is introduced by real-time deployment. Although they may boost representation capacity, larger transformers may also result in higher latency and memory usage. Thus, possible optimisation techniques include edge inference, frame skipping, quantisation, pruning, and asynchronous explanation generation. Deployability and networked edge/cloud systems are identified as key study areas in recent VAD literature.

Another significant issue is generalisation between cities. The distribution of data can be altered by various camera heights, weather, lighting, traffic patterns, and social behaviours. As a result, a random frame split may exaggerate actual performance. Before declaring deployment readiness, cross-camera and cross-location testing have to be incorporated.

10. Limitations

The camera network and emergency response process of a particular city cannot be fully replicated using public datasets. Additionally, rare events can be under-represented. Predictions and

explanations might be impacted by dataset bias. Furthermore, XAI techniques offer attribution as opposed to causal explanations. Lastly, without site-specific validation, threshold calibration, fail-safe procedures, and human oversight, a model should not be directly linked to safety-critical automated activities.

11. Conclusion and Future Work

An explainable spatial-temporal framework for real-time critical-event detection and prediction in smart cities was presented in this paper: XAI-CityVision. Heterogeneous sensing, computer vision, object detection, temporal modelling, risk assessment, and explainable AI are all integrated into the system. Predictive accuracy, false alarms, latency, throughput, ablation behaviour, and explanation quality are all assessed using the suggested experimental technique. Continuous learning, privacy-preserving edge intelligence, multi-camera event association, vision-language models, uncertainty-aware alerting, and cross-city domain adaptation will all be the subject of future research.

12. References

1. Samaila, Y. A., et al. “Video anomaly detection: A systematic review of issues and prospects.” Neurocomputing, 591, 127726, 2024. DOI: 10.1016/j.neucom.2024.127726.
2. “Deep Learning for Video Anomaly Detection: A Review.” IEEE Transactions on Neural Networks and Learning Systems, 37(7), 3010–3030, 2026. DOI: 10.1109/TNNLS.2025.3647892.
3. “Video Surveillance and Artificial Intelligence for Urban Security in Smart Cities: A Review of a Selection of Empirical Studies from 2018 to 2024.” Smart Cities, 10(1), 15, 2025.
4. Sharif, M. H., Jiao, L., & Omlin, C. W. “Deep crowd anomaly detection: state-of-the-art, challenges, and future research directions.” Artificial Intelligence Review, 58, 139, 2025. DOI: 10.1007/s10462-024-11092-8.
5. Liu, J., Liu, Y., Lin, J., et al. “Networking Systems for Video Anomaly Detection: A Tutorial and Survey.” ACM Computing Surveys, 57(10), 2025. DOI: 10.1145/3729222.
6. “Survey on video anomaly detection in dynamic scenes with moving cameras.” Artificial Intelligence Review. DOI: 10.1007/s10462-023-10609-x.
7. Wang, Y., Guo, D., Li, S., Camps, O., & Fu, Y. “Explainable Anomaly Detection in Images and Videos: A Survey.” arXiv:2302.06670, 2023.
8. Ren, Y., et al. “A Panoramic Review on Cutting-Edge Methods for Video Anomaly Localization.” IEEE Access, 12, 186380–186412, 2024. DOI: 10.1109/ACCESS.2024.3510039.
9. Sivakumar, G., Mogesh, G., Pragatheeswaran, N., & Sambathkumar, T. “Video Anomaly Detection in Crime Analysis Using Deep Learning Architecture—A Survey.” Journal of Trends in Computer Science and Smart Technology, 6(1), 1–17, 2024. DOI: 10.36548/jtcsst.2024.1.001.
10. Baala, A., Mostafa, H., & Mohssine, B. “A Comprehensive Systematic Review of Deep Learning Techniques for Anomaly Detection in Urban Video Surveillance.” IEEE IRASET, 2025. DOI: 10.1109/IRASET64571.2025.11008153.
11. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. “You Only Look Once: Unified, Real-Time Object Detection.” CVPR, 779–788, 2016.
12. Dr. J. Anvar Shathik *et al.* ” Deep Ensemble Learning For Accurate Prediction Of Neurodegenerative Disorders Using Temporal Clinical Data . (2025). *International Journal of Environmental Sciences*, *11*(4s), 1246-1253. https://doi.org/10.64252/8dbg7v71
13. Anvar Shathik J, B. R. Chandra, M. D. Nandeesh, T. C. Maniunath, S. Pothala and N. R. Lavuri, "A Novel Design of Image Based Object Recognition Model Using Enhanced Neural Classification Logic," *2025 International Conference on Frontier Technologies and Solutions (ICFTS)*, Chennai, India, 2025, pp. 1-8, doi: 10.1109/ICFTS62006.2025.11031570.
14. R. R, J. Anvar. Shathik, J. Sivakumar, P. Manothini, G. S. J. Asha and M. Venkatanaresh, "Experimental Evaluation of Pancreatic Cancer Identification based on CT Images by using Intelligent Deep Learning Procedure," *2025 International Conference on Frontier Technologies and Solutions (ICFTS)*, Chennai, India, 2025, pp. 1-9, doi: 10.1109/ICFTS62006.2025.11031801.
15. J. Anvar Shathik, S. Hashini, A. Pandiaraj, N. Ramshankar, N. G and A. K B, "Identification of Different Medicinal Plants Through Image Processing," *2024 International Conference on Smart*

Technologies for Sustainable Development Goals (ICSTSDG), Chennai - 600077, Tamil Nadu, India, 2024, pp. 1-5, doi: 10.1109/ICSTSDG61998.2024.11026565.

16. Raju, K., Ramshankar, N., Anvar Shathik J *et al.* Blockchain Assisted Cloud Security and Privacy Preservation using Hybridized Encryption and Deep Learning Mechanism in IoT-Healthcare Application. *J Grid Computing* **21**, 45 (2023). https://doi.org/10.1007/s10723-023-09678-7
17. Madhavikatamaneni, R. K. S, Anvar Shathik J and K. PoornaPushkala, "A Healthcare System for detecting Stress from ECG signals and improving the human emotional," *2022 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, Coimbatore, India, 2022, pp. 1-8, doi: 10.1109/ICACTA54488.2022.9753564.
18. VijayaVardan Reddy S P; Armstrong Joseph J; Priscilla M; Anvar Shathik J; R.ThandaiahPrabu, "HDP-IoT: An IoT Framework for Cardiac Status Prediction System using Machine Learning," *2022 International Conference on Inventive Computation Technologies (ICICT)*, Nepal, 2022, pp. 855-861, doi: 10.1109/ICICT54344.2022.9850897.
19. K. Vikranth, Anvar Shathik J, and K. Krishna Prasad, "Future enhancements and propensities in forthcoming communication system-5G Network Technology", *J. Phys*, pp. 12006, 2020.
20. Yarramsetti, S., Anvar Shathik J, & Renisha, P. S. (2021). Intelligent estimation of social media sentimental features using Deep Learning with Natural Language Processing Strategies. *International Journal of Innovative Technology and Exploring Engineering*, *10*(6), 74-79.
21. Ramshankar, N., Anvar Shathik, J.., Raju, K. *et al.* Integrated deep learning and blockchain-based framework for cloud manufacturing with improved customer satisfaction. *KnowlInfSyst* **67**, 5301–5334 (2025). https://doi.org/10.1007/s10115-025-02373-x
22. Mr.G. Silambarasan , J.Anvar Shathik (2017) "Ensemble text classifier: a document classification technique to predict and categorizes regularised and novel classes using incremental learning " , International Journal of Applied Engineering Research , 2017, Volume 12, issue -22 , Pages12454-12459
23. Mr.G. Silambarasan , J. Anvar Shathik (2019) " Automated Ensemble Framework for Integration of Ontology Based Large Scale Semantic Knowledge Base " Journal of Engineering and Applied Sciences 14(2)
24. J. Anvar Shathik, Anil Saroliya,*et al*.(2025) "Smart Vision Systems for Public Safety: Real-Time Crowd Monitoring and Anomaly Detection in Urban Spaces Using Deep Learning and Edge Computing", International Journal of Applied Mathematics, 38 (6) **DOI:** https://doi.org/10.12732/ijam.v38i6s.430
25. Anvar Shathik J; , B. R. Chandra, M. D. Nandeesh, T. C. Maniunath, S. Pothala and N. R. Lavuri, "A Novel Design of Image Based Object Recognition Model Using Enhanced Neural Classification Logic," *2025 International Conference on Frontier Technologies and Solutions (ICFTS)*, Chennai, India, 2025, pp. 1-8, doi: 10.1109/ICFTS62006.2025.11031570.