

Research Paper

Multiscale Deep Learning Framework for Human Behaviour Classification and Analysis

Zoya Tabassum¹, Auof Ali Khan², Mohammed Ateeq³, Mahey Zia⁴, Syed Abdul Rahman Qadri⁵

¹Department of CSE (Data Science),

Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.

^{2,3,4,5}UG Students, Department of CSE (Data Science),

Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.

Abstract— Human behaviour identification is a significant research domain in computer vision and deep learning, as it has extensive use in areas such as observation, healthcare monitoring, and intelligent systems. Nevertheless, precisely discerning human actions from visual data remains difficult due to fluctuations in mobility, diverse camera angles, and evolving settings. Numerous current systems employ conventional machine learning techniques or rudimentary convolutional neural networks, that frequently fail to adequately capture both temporal and spatial information. This research presents a multi-scale deep learning techniques framework for the classification and analysis of human behaviour. The system initiates with dataset upload, preprocessing, and partitioning to ready the data for testing and training purposes. The basic CNN2D model is employed to extract spatial data from images, whilst a CNN3D model is implemented to collect spatial as well as temporal characteristics from video sequences. A hybrid model integrating CNN, GRU, and bidirectional layers is generated to enhance the comprehension of sequential patterns and elevate overall predictive performance. The system comprises modules for showing accuracy and loss graphs, in addition to facilitating the comparison of various models' performance. Experimental findings

indicate that the suggested framework attains superior accuracy and enhanced performance relative to conventional methods. This enhances the system's reliability for real-time human behaviour identification tasks.

Keywords— Human Behaviour Recognition, Deep Learning, CNN2D, CNN3D, GRU, Bidirectional Networks, Multiscale Framework, Activity Classification

I. INTRODUCTION

Human behavior recognition (HBR) represents a significant and expanding domain within computer vision. It is extensively utilized in practical applications like surveillance systems, healthcare observation, human-computer interaction, and intelligent settings. The primary objective of HBR research is to autonomously recognize and categorize human actions using video or data from sensors. Nonetheless, this task is challenging due to the variability of human movements influenced by variances in mobility, angles of camera, background, lighting circumstances, and occlusions, as elaborated in [1] and [4]. These issues get increasingly intricate in real-world contexts, as emphasized in [6].

Previous methodologies for behaviour recognition predominantly depended on conventional techniques, including the extraction of features, feature descriptions, and classification. These methodologies

employ techniques such as detection of edges, motion tracking, and manually built characteristics to identify human actions. While they function to a degree, they exhibit sluggishness, less adaptability, and significant susceptibility to noise and environmental fluctuations, hence constraining their efficacy in real-world applications as detailed in [2]. Comparable challenges have been documented in [8] as well as [12], wherein comparable approaches exhibit worse performance in intricate situations.

The advancement of deep learning, particularly convolutional neural networks (CNNs), has markedly enhanced the efficacy of behavior recognition systems. These models can autonomously extract features from data, diminishing the necessity for manual labor. Models are categorized into 2-dimensional CNN & 3D CNN based on the application of convolution. The 2D CNN primarily emphasizes spatial characteristics in images, although it is unable of capturing temporal motion. Conversely, the 3D CNN is capable of processing video sequences and assimilating both time and space information, rendering it more proficient at recognizing dynamic human actions, as demonstrated in [5]. An enhanced temporal modeling method has been presented in [9], which augments achievements in real-time applications.

In recent years, numerous sophisticated models, including C3D, ECO, and TSN, have been created to enhance behavior recognition. These models can acquire superior feature representations and differentiate between various activities with greater precision. Researchers have additionally focused on enhancing performance using tactics such as attention processes and optimization methodologies, as outlined in [10]. The significance of deep learning in enhancing behavior recognition has been emphasized in [13] and [18], resulting in improved accuracy and robustness.

Despite these enhancements, difficulties persist. Human behavior is intricate and may vary according to differing circumstances. Single-scale models fail to capture characteristics at varying levels, hence diminishing accuracy. Additionally, issues such as occlusion, ambient noise, and minor or sluggish movements complicate recognition, as noted in [3], along with [7]. Environmental variables and system constraints also influence performance, as elaborated in [11].

This research presents a multi-scale deep learning framework that allows the classification and analysis

of human behavior to overcome these difficulties. The proposed system integrates many models, including CNN2D for spatial feature extraction, CNN3D for spatiotemporal information capture, and hybrid models such as CNN , GRU with bidirectional layer for sequence learning. This methodology enhances feature learning, elevates model performance, and attains superior accuracy relative to current techniques.

II. LITERATURE SURVEY

The recognition of human behavior through skeletal data has garnered considerable interest due to its clear and shortened feature representation, which is less susceptible to fluctuations in appearance. In the research conducted by [M. Z. Sun] et al. in [2022] [2], a refined feature extraction technique named ExGist was introduced to enhance the representation of human bone joint data for action recognition purposes. The approach use OpenPose to obtain joint coordinates and integrates it into thermodynamic mapping to improve feature representation. The ExGist Feature feature is intended to encapsulate local and global attributes of skeletal information while preserving the benefits of the conventional Gist feature. Experimental findings indicate that ExGist outperforms the conventional Gist feature across many classifiers. Moreover, in comparison to advanced models like C3D as well as APTNet, the proposed approach attained superior accuracy, achieving up to 89.2%, hence illustrating its efficacy in enhancing human behavior recognition performance.

The burgeoning population is amplifying demand for effective and high-quality healthcare systems, so rendering patient monitoring a crucial component of contemporary medical care. The study by Z. He (2022) proposes a system based on machine learning designed to monitor patient behaviors and identify potential health risks through the use of many cameras positioned at various angles to record extensive behavioral data. The system evaluates physiological and behavioral patterns to detect potential medical issues, emphasizing high accuracy and performance in real time, hence enhancing

patient care and alleviating the work load of healthcare personnel. Two prominent data set, ETRI-Activity3D along with NTU-RGB-D, are employed for evaluation; the ETRI dataset encapsulates everyday patient activities, while the NTU-RGB-D dataset emphasizes medical condition identification. The data undergoes pre-processing to eliminate extraneous information and decrease computational complexity, subsequently employing a multi-feature descriptor method to extract significant features. The features are enhanced with matching features and optimization techniques, resulting in a classification model that attains an accuracy with 87.5% on the ETRI dataset as well 89.5% with the NTU-RGB-D dataset, thereby illustrating the efficacy of the proposed method in practical healthcare applications.

Currently, surveillance footage systems are crucial for detecting hazardous situations involving pedestrians. In the research proposed by [C. Y. Zhang] et al. [2022] [4], an approach is established to autonomously identify individuals and interpret their behavior from video recordings. The system employs this YOLO (You Only Look Once) technique to identify pedestrians in live video and monitor their movements by evaluating both current and prior frames. Consequently, it can recognize various motions, like ambulation, sprinting, or remaining still, and can even discern atypical or dubious movement patterns. The findings indicate that the system functions efficiently in identifying both normal and deviant behaviors. The system processes up to 77 pixels per second with YOLOv3-tiny & 21 frame per second with YOLOv3, rendering it appropriate for live monitoring when utilized with a graphically supported GPU system. superior diagnosis accuracy compared to manual approaches, enhancing interpretability and application in healthcare environments, signifying improvements in precision and decrease of errors.

In video-based human action detection, two-stream networks are extensively utilized due to their capacity to independently analyze spatial and temporal information; nevertheless, several existing methodologies employ identical network architectures for both streams, hence constraining their efficacy. To address this issue, [X. Ding] et al. [2022] [5] introduced a spatiotemporal homogeneous two-stream network that employs distinct network structures for spatial and temporal extraction of features, namely ResNet and BN-Inception, to more effectively capture the nuances of human activities.

The model employs a segment-based methodology to comprehend long-term temporal linkages in video sequences, facilitating the differentiation of actions with analogous sub-actions. Furthermore, data augmentation and an enhanced cross-modal pre-training method are employed to boost performance. Experimental findings on benchmark datasets, including UCF101 and HMDB51, indicate that proposed method outperforms conventional two-stream structures and other current methodologies.

The research conducted by S. Huang in 2021[6] provides a comprehensive examination of behavior detection technology and its potential applications within the realm of computer vision. The paper elucidates fundamental principles and various methodologies employed in behavior recognition, encompassing strategies reliant on multimodal information, depth images, skeleton key features, and RGB data, allowing for precise analysis of human motions. This also underscores the extensive applications of this type of technology in domains like safety and education. In safety-related contexts, behavior recognition can identify atypical or hazardous activities and enhance security protocols, whilst in education, it can aid educators in comprehending student performance and refining instructional strategies. The device can also discern behavioral patterns that signify when a pupil may require supplementary assistance. The report additionally explores prospective research avenues, including the creation of more precise and efficient models, as well as tackling significant issues pertaining to ethics and privacy. The study highlights the significant potential of behavior recognition technology and advocates for additional research to enhance its applications and efficacy.

III. DATASET DESCRIPTION

In our proposed study, we utilized a Human Activity Recognition (HAR) datasets comprising data gathered from smartphone sensors to recognize various human activities. The dataset include actions like ambulation, sitting, standing, reclining, and various other quotidian movements. The dataset comprises records that include sensor values paired with appropriate activity labels, facilitating the model's ability to learn and properly categorize human behavior. Data is collected by sensors such as an accelerometer and the gyroscope, which quantify body movement, velocity, and orientation. The sensor readings are the primary attributes of a dataset and serve as input data for the model.

Upon loading the dataset into an application, it is presented and analyzed to comprehend the distribution of activities. The dataset undergoes preparation stages including cleaning, normalization, and elimination of extraneous noise. Graphs are constructed to illustrate various activities, with the x-axis denoting activity titles and the y-axis indicating the frequency of each action. Subsequent to the preprocessing phase, the dataset is partitioned into sets for testing and training for model development. The organized and refined dataset is subsequently utilized to train various models, including CNN2D, CNN3D, and CNN integrated with GRU and Bidirectional layers, facilitating precise predictions of human behaviors.

IV. IMPLEMENTATION DETAILS

The execution of the suggested system adheres to a systematic workflow, commencing with dataset loading and culminating in final prediction. The Human Activity Recognition (HAR) dataset is initially uploaded into the application via the Tkinter-based interface. Upon loading the dataset, it is presented for the user's comprehension of its structure as well as contents. The dataset comprises sensor-derived information obtained from smartphones, reflecting various human activities. Prior to utilizing the data to training, preprocessing operation is conducted to cleanse the dataset by eliminating noise, addressing values that are missing, and normalization the features to ensure all values reside within a comparable range.

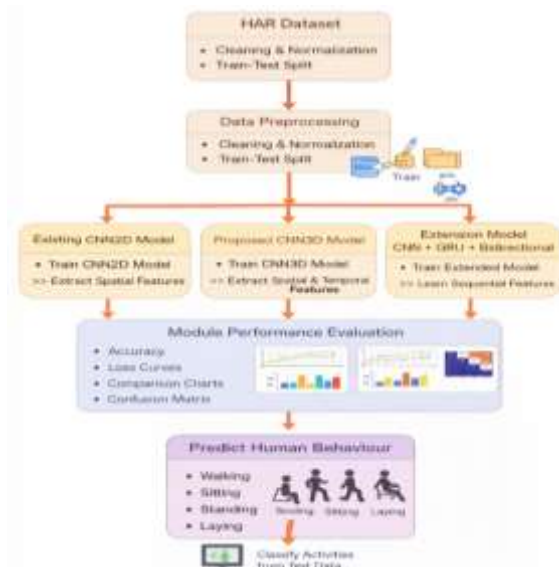
Subsequent to preprocessing, the dataset undergoes analysis and visualization to comprehend the distribution of various activities. Graphs are constructed using activity names displayed on the horizontal axis and their respective counts on the y axes. This assists in determining the balance of the dataset. The dataset is subsequently partitioned into two segments: training data as well as testing data, respectively. Training dataset is utilized in order to develop the models, whereas a testing dataset is employed to assess their performance on novel data.

Initially, a pre-existing CNN2D model is employed as a baseline methodology. This model emphasizes the extraction of spatial information from the dataset & the acquisition of patterns associated with various activities. This model necessitates a greater number of parameters, hence augmenting computing complexity and training duration. Despite attaining

an accuracy of approximately 92%, it exhibits problems in extracting temporally information derived from sequential data.

The proposed CNN3D (MCNN/MDN) architecture is provided to enhance performance. This model is engineered to encapsulate both temporal and spatial characteristics, enhancing its efficacy in human activity recognition. In comparison to CNN2D, it utilizes less parameters and attains superior accuracy of approximately 94%. This model offers a more nuanced comprehension of human behavior by analyzing motion patterns throughout time.

An enhanced extension model including CNN, GRU, & Bidirectional layer has been designed to significantly improve performance. This model employs CNN for extracting features, whereas GRU & Bidirectional layers facilitate the acquisition of sequence and contextual information derived from the data. This combination decreases the parameter count and enhances efficiency while attaining a maximum accuracy of approximately 96%. The system incorporates performance assessment through accuracy graphs and loss curves, comparative charts, & confusion matrices. Furthermore, it facilitates real-time prediction, allowing new input data to be sent to the model for the classification of actions such as



moving, sitting, and standing.

Figure 1: System Architecture for Proposed Multiscale Deep Learning Framework for Human Activity Recognition

V. RESULT AND DISCUSSION

The initial phase of our suggested system commences with executing the application via the run file, which



activates the Tkinter-based interface. This interface offers several functionalities, including dataset submission, preprocessing, model execution, and human behavior prediction. Upon pressing given "Upload Dataset" button, and the system processes the dataset and presents its values, thereby confirming the successful upload of the dataset.

Figure [2] illustrates that the interface enables users to do additional tasks such as preprocessing, training models, and visualization. This screen serves as the first interface of the system, enabling user interaction with the application as well as management of the entire workflow.

Figure 2: User Interface for Dataset Upload and Model Execution

Upon training the current CNN2D model, the system presents its performance metrics & confusion matrix. Figure [3] illustrates that the model attains approximately 92% accuracy, alongside precision, recall, & F1-score. The confusion matrix displays expected labels along the x-axis & actual labels along the y-axis, facilitating the identification of accurate and erroneous predictions. The model demonstrates commendable performance; yet, it exhibits several limits when juxtaposed with more advanced versions.



Figure 3: Performance Evaluation and Confusion Matrix of Existing CNN2D Model

Matrix of Existing CNN2D Model

Upon training the recommended CNN3D (MDN) method, the platform presents metrics for performance and confusion matrix. Figure [4] illustrates that the model attains superior accuracy compared to the CNN2D model, along with enhanced precision, recall, as well as F1-score. The confusion matrix displays anticipated labels along the x-axis along with actual labels along the y-axis, with the majority of values positioned on the diagonal, signifying accurate predictions. The CNN3D model demonstrates superior efficacy in capturing all spatial and temporal data.

Figure 4: Performance Evaluation and Confusion Matrix of Proposed CNN3D (MDN) Model

Upon training an extension model, the model presents its metrics for performance and confusion



matrix. Figure [5] illustrates that the model attains superior accuracy, alongside enhanced precision, recall, and F1-score, in comparison to other models. The confusion matrix displays anticipated labels on the axis of x and genuine labels over the y-axis, with the majority of values positioned along the diagonal, signifying accurate predictions. The extension model demonstrates superior effectiveness in the classification of human activities.



Figure 5: Performance Evaluation and Confusion Matrix of Extension Model

Upon completion of model training, the system presents a comparative graph illustrating the performance for each algorithm. Figure [6] illustrates a comparison of metrics including precision, recall, accuracy, and F1-score across various models. The x-axis denotes the algorithms, and the y-axis indicates the performance numbers. The graph demonstrates that an extension model outperforms the other models, signifying its superior efficacy in human behavior classification.

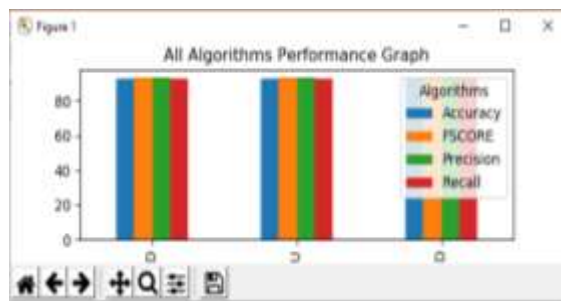


Figure 6: Comparison Graph of All Models Performance

DISCUSSION

In a recent period, human behaviour recognition has emerged as a significant study domain owing to its extensive applications in the healthcare sector, surveillance, and intelligent systems. The primary objective of this study was to create an efficient and precise model for identifying human activities via deep learning methodologies. Conventional techniques and simplistic models often fail to accurately represent intricate movement patterns, particularly in the context of temporal data. This work employed the Human Activity Recognition (HAR) dataset of sensor-based features to train and assess various models. The CNN2D model was initially employed, yielding satisfactory performance but exhibiting constraints in managing temporal dependencies. A CNN3D architecture was

implemented in order to enhance performance by effectively capturing both temporal and spatial data, resulting in improved accuracy. An advanced model integrating CNN, GRU, with Bidirectional layers was created to improve sequential learning and contextual comprehension. This hybrid methodology enhanced the model's performance by enabling the acquisition of more intricate patterns within the data. The findings indicate the extension model surpasses other models regarding accuracy and reliability. The implementation of multiscale deep learning methods enhances the system's robustness, rendering it appropriate for real-world human behavior identification applications.

VI. CONCLUSION

This project aims to enhance human behavior recognition through deep learning methodologies. The suggested system illustrates how the integration of several models can improve classification efficacy and address the shortcomings of conventional methods. This study implemented and compared various models, including CNN2D, CNN3D, as well as a hybrid model utilizing CNN, GRU, with Bidirectional layers. The findings indicate that the enhanced model delivers superior accuracy and more dependable predictions by adeptly capturing combined spatial and temporal characteristics. The system enhances learning efficiency with model performance through the utilization of a structured dataset and the implementation of appropriate preprocessing procedures. This methodology underscores the significance of integrating various deep learning techniques to develop a robust and precise predictive system. The proposed model enhances the advancement of effective human behavior detection systems and is applicable in real-world contexts including healthcare monitoring, surveillance, even intelligent surroundings.

REFERENCES

1. X.-J. Gu, P. Shen, H.-W. Liu, J. Guo and Z.-F. Wei, "Human behavior recognition based on bone spatio-temporal map", *Comput. Eng. Des.*, vol. 43, no. 4, pp. 1166-1172, 2022.
 - a. Show in Context [Google Scholar](#)
2. M. Z. Sun, P. Zhang and B. Su, "Overview of human behavior recognition methods based on bone data features", *Softw. Guide*, vol. 21, no. 4, pp. 233-239, 2022.

- a. Show in Context [Google Scholar](#)
3. Z. He, "Design and implementation of rehabilitation evaluation system for the disabled based on behavior recognition", *J. Changsha Civil Affairs Vocational Tech. College*, vol. 29, no. 1, pp. 134-136, 2022.
 - a. Show in Context [Google Scholar](#)
4. C. Y. Zhang, H. Zhang, W. He, F. Zhao, W. Q. Li, T. Y. Xu, et al., "Video based pedestrian detection and behavior recognition", *China Sci. Technol. Inf.*, vol. 11, no. 6, pp. 132-135, 2022.
 - a. Show in Context [Google Scholar](#)
5. X. Ding, Y. Zhu, H. Zhu and G. Liu, "Behavior recognition based on spatiotemporal heterogeneous two stream convolution network", *Comput. Appl. Softw.*, vol. 39, no. 3, pp. 154-158, 2022.
 - a. Show in Context [Google Scholar](#)
6. S. Huang, "Progress and application prospect of video behavior recognition", *High Tech Ind.*, vol. 27, no. 12, pp. 38-41, 2021.
 - a. Show in Context [Google Scholar](#)
7. Y. Lu, L. Fan, L. Guo, L. Qiu and Y. Lu, "Identification method and experiment of unsafe behaviors of subway passengers based on Kinect", *China Work Saf. Sci. Technol.*, vol. 17, no. 12, pp. 162-168, 2021.
 - a. Show in Context [Google Scholar](#)
8. X. Ma and J. Li, "Interactive behavior recognition based on low rank sparse optimization", *J. Inner Mongolia Univ. Sci. Technol.*, vol. 40, no. 4, pp. 375-381, 2021.
 - a. Show in Context [Google Scholar](#)
9. Z. Zhai and Y. Zhao, "DS convLSTM: A lightweight video behavior recognition model for edge environment", *J. Commun. Univ. China Natural Science Ed.*, vol. 28, no. 6, pp. 17-22, 2021.
 - a. Show in Context [Google Scholar](#)
10. C. Ying and S. Gong, "Human behavior recognition network based on improved channel attention mechanism", *J. Electron. Inf.*, vol. 43, no. 12, pp. 3538-3545, 2021.
 - a. Show in Context [Google Scholar](#)
11. Z. Duan, Q. Ding, J. Wang and W. Li, "Subway station lighting control method based on passenger behavior recognition", *J. Railway Sci. Eng.*, vol. 18, no. 12, pp. 3138-3145, 2021.
 - a. Show in Context [Google Scholar](#)
12. D. Liu, J. Yang and Q. Tang, "Research on identification technology of violations in key underground places based on video analysis", *Proc. Excellent Papers Annu. Meeting Chongqing Mining Soc.*, pp. 71-75, 2021.
 - a. Show in Context [Google Scholar](#)
13. Y. Ye, "Key technology of human behavior recognition in intelligent device forensics based on deep learning", Apr. 2021.
 - a. Show in Context [Google Scholar](#)
14. Y. Li, "Mining the spatiotemporal distribution law of CNG gas dispensing sub station and identifying abnormal behaviors based on machine learning", Apr. 2021.
 - a. Show in Context [Google Scholar](#)
15. W. Wang, "Research on behavior recognition based on video image and virtual reality interaction application", Mar. 2021.
 - a. Show in Context [Google Scholar](#)
16. J. Wang, "Design and implementation of enterprise e-mail security analysis platform based on user behavior identification", *J. Shanghai Inst. Shipping Transp. Sci.*, vol. 43, no. 4, pp. 59-64, 2020.
 - a. Show in Context [Google Scholar](#)
17. K. Han and Z. Huang, "A fall behavior recognition method based on the dynamic characteristics of human posture", *J. Hunan*

- Univ. Natural Sci. Ed., vol. 47, no. 12, pp. 69-76, 2020.
 - a. Show in Context [Google Scholar](#)
18. F. Wang, "Research on attitude estimation and behavior recognition based on deep learning in logistics warehousing", 2020.
 - a. Show in Context [Google Scholar](#)
19. Y. Ying, "Analysis of prenatal behavior characteristics of Hu sheep and development of monitoring system based on embedded system", 2020.
 - a. Show in Context [Google Scholar](#)
20. J. Bao and H. Jin, "A semi supervised learning method for identifying intrusion behaviors in ship LAN", *Ship Sci. Technol.*, vol. 42, no. 24, pp. 136-138, 2020.
21. Show in Context [Google Scholar](#)
22. Babburi, S. Privacy-Preserving Collaborative Framework with Auditable Federated Learning.
23. Gaddam, S. Integrating Analytics into the Development Process: Bridging the Gap between Data Insights and Design Execution.
24. Immadi, S. K. (2025). Optimizing ERP for Human Capital Management. *Applied Research for Growth, Innovation and Sustainable Impact*, 377–384. <https://doi.org/10.1201/9781003684657-63>
25. Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.
26. Poojari, R. INTELLIGENT SYSTEMS+B108 AND APPLICATIONS IN ENGINEERING.
27. Mahimalur, R. K., Vasgam, M., & Manoharan, D. Devops Lifecycle Management And Cloud Migration Assessments: A Security-Driven CICD Perspective.
28. Viswanathan, V. (2023). AI-Augmented Decision Intelligence for Enterprise Systems: Integrating Cognitive Analytics for Resource and Talent Optimization.
29. Agrawal, A. M., Gajula, S., Shinde, R. P., Shah, H., & Ghosh, H. (2025, July). Machine Translation for Long Sequences with Enhanced Attention Mechanisms. In 2025 5th International Conference on Electrical, Computer and Energy Technologies (ICECET) (pp. 1-6). IEEE.
30. Maturi, S. Y. (2021). Blockbond hardening: Securing pooled-hash protocols against traffic tampering, MITM hash-rate hijacking, and template coercion. *International Journal of Communication Networks and Information Security*, 13(3), 718–728.
31. Adabala, P. K. (2024). Utilizing predictive analytics to improve efficiency and decision-making in ERP-connected supply chains. *International Journal of Intelligent Systems and Applications in Engineering*, 12(22s), 2465
32. Kavuri, S. (2026). An Explainable Machine Learning Framework for Predicting Software Defects in Large-Scale Software Systems. 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC), 1–6. <https://doi.org/10.1109/icaic67076.2026.11395777>
33. Gajula, S. (2025). Ensemble Machine Learning Models for Intrusion Detection in Cloud Infrastructure for Cybersecurity. 2025 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD), 1–6. <https://doi.org/10.1109/icoabcd67551.2025.11470865>
34. Shashank, A. (2025). Self-Healing Data Pipelines for Enhanced Reliability: A Paradigm Shift in Enterprise Data Management. *Journal of Computer Science and Technology Studies*, 7(8), 1097-1104.
35. Susarla, R. S., Boyapati, P. K., & Kandula, S. T. R. (2025, July). Cloud-Based Secure Data Storage in Smart Cities Using Central-Smoothing Hypergraph Neural Networks. In 2025 IEEE 4th World Conference on Applied Intelligence and Computing (AIC) (pp. 279-284). IEEE.

36. Boyapati, P. K. Building a centralized data operations hub for healthcare enterprise integration. IJSAT-Int. J. Sci. Technol. 16 (2).
<https://doi.org/10.71097/IJSAT.v16.i2.3219>