International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

AI-Based Surveillance Framework for Physical Violence Detection

Srividya M. S.*[1], Dr. Anala M. R.[2]

Submitted: 25/01/2024 **Revised:** 03/03/2024 **Accepted:** 11/03/2024

Abstract: This research paper presents a groundbreaking approach to addressing the societal problem of physical abuse, which affects various demographic groups, including children, women, and older people, especially in domestic and workplace environments. The complexity of these situations, especially when the abuser and victim know each other, highlights the need for an advanced solution. The paper introduces a novel hybrid deep-learning framework to detect and prevent physical abuse and address this. The framework utilizes human action recognition, leveraging a 3D convolutional neural network (CNN) to meticulously analyze human actions in such contexts. The deep learning model is further enhanced by employing transfer learning techniques with ResNet-18 and GoogleNet models. These models are trained using the UBI fight and UCF crime datasets, which are public resources for video analysis, to identify instances of physical abuse. A significant innovation in this model is the transformation of 2D kernels into 3D kernels, which allows for an improved extraction of features in both temporal and spatial dimensions from the video data. Additionally, a bilinear Long Short-Term Memory (LSTM) layer is integrated into the model to capture more extended material information, thus improving the analysis of human actions. The results of this hybrid model in detecting physical abuse are promising, showing marked improvements in performance metrics due to the shift from 2D to 3D kernels and the inclusion of bilinear LSTM.

Keywords: *Deep Learning, Human Action Recognition, 3D Convolutional Neural Networks, Long Short-Term Memory (LSTM), Physical Abuse Detection, Transfer Learning, Video Surveillance Analysis, Performance Metrics in Machine Learning.*

1. Introduction

Throughout history, the expression of anger, an emotion deeply rooted in human experience, has often manifested in physical ways across diverse cultures. Unfortunately, this tendency to express anger physically results in harm, particularly to vulnerable groups. This includes severe injuries and lasting psychological impacts. According to World Health Organization data, instances of physical abuse directed at women and children are notably prevalent in countries like the United States and India. The issue of violence extends to educational settings such as schools and colleges, as well as other institutions like childcare centers and hostels. While surveillance systems, often through CCTV, can detect such incidents, they require human oversight for effective intervention. This necessitates the development of more sophisticated, intelligent surveillance technologies.

The challenge lies in accurately identifying human behaviors, an intricate task that has become a key area of focus in fields like Machine Learning and Computer Vision. Algorithms rooted in computer vision [1] face challenges in accurately capturing the nuances of complex human actions within a given scene. Recognizing these actions involves more than just analyzing the movement patterns of body parts [2]; it also requires understanding the context, cultural factors, and the collective behavior of all individuals present. This makes the process of extracting relevant features particularly complex [3].

Deep Learning (DL) has emerged as the most effective method for tackling this complexity. The evolution of GPU architecture and the vast amounts of video data generated by the rise of social media [4] have significantly advanced DL's capabilities. Additionally, the development of depth-sensing camera technologies has been revolutionary, enabling the capture of detailed 3D structures and postures of the human body, thus offering a more comprehensive view of human actions.

The advancements in artificial intelligence, particularly with Deep Learning (DL) algorithms, have been transformative, propelling Convolutional Neural Networks (CNNs) to the forefront as robust tools for tasks like feature extraction and classification [5]. CNNs are distinguished by their ability to autonomously learn high-level features from raw data, drastically reducing the need for complex manual feature engineering. This capability is a significant breakthrough, streamlining processes that were once time-consuming and expertise-dependent. CNNs, inspired by the human visual cortex, excel in interpreting visual data, making them highly effective in various computer vision domains. Their architecture, consisting of layers of convolutional filters, is adept at identifying patterns, edges, and shapes in images.

This proficiency extends beyond visual tasks, finding

1Assistant Professor, Department of Computer Science and Engineering,
R V College of Engineering, Bengaluru-59, India

2Professor, Department of Information Science and Engineering, R V College of Engineering, Bengaluru-59, India
** Corresponding Author Email: srividyams.publications@gmail.com*

applications in natural language processing and medical image analysis. In healthcare, for instance, CNNs are revolutionizing medical diagnostics by analyzing images such as X-rays and MRIs with remarkable accuracy. The widespread application of CNNs across these diverse fields marks a substantial technological leap, enhancing our ability to understand and respond to complex visual information and human actions and opening up new avenues for innovation and application.

2. Literature Survey

The study of human action recognition is increasingly prominent in computer vision research. The field, known as Human Activity Recognition (HAR), aims to identify and categorize human activities like jumping, playing, punching, walking, and running in various forms of media such as images and videos. Vision-based methods are particularly noted for their effectiveness in capturing temporal actions, a point emphasized by Yanmin [6]. Yet, the consistent and accurate recognition of human behavior in video format poses a substantial challenge.

Recent advancements include a novel algorithm by Waheed [7], achieving an impressive 98% accuracy on test samples. This result was obtained by training with a specified batch size and learning rate, employing the Adam optimizer. HAR now incorporates various data types, from skeleton and point cloud data to infrared, depth, and even radar signals, as explored in Zehua's work [8]. Thomas [9] found particular strengths in audio data for temporal sequence localization and acceleration data for refining HAR, while radar data has enabled HAR even through walls.

Ganesh [10] introduced the concept of Skepxels, utilizing CNNs to create intricate connections between skeletal joints. This approach is complemented by research in robotic engineering, which incorporates diverse human skeletal datasets [11]. Kim [12] tailored a strategy for virtual sports training, focusing on precise 3D skeleton data collection applied to sports like boxing and tennis. Nouray's survey [13] discusses various technologies in sports-related HAR.

Neziha et al. [14] proposed a hybrid DL model for recognizing human behavior, emphasizing the significance of feature classification. This model was tested on the UCF sports action dataset, and the Gaussian Mixture Model was combined with the Kalman Filter and GRU, achieving 96.3% accuracy on the KTH dataset. There has also been research on learning interaction patterns directly from video data [15], with recent studies employing CNN-based techniques [16-20] to analyze individual poses and postures.

In medical imaging, Raza et al. [21] developed DeepTumorNet, a CNN model for brain tumor classification, using a comprehensive dataset of CE-MRI scans. Similarly, Ritu Tandon et al. [22] introduced a hybrid DL model, VCNet, for lung cancer nodule detection in CT scans using various pre-trained CNN models.

Hnamte [23] proposed a novel two-stage DL model combining Auto-Encoders and LSTM for attack detection in cyber security, showing promising results. Similarly, Khatan et al. [29] developed an effective Intrusion Detection System using CNN and DNN, comparing their model's effectiveness with others in the field.

In a series of studies exploring the applications of LSTM (Long Short-Term Memory) networks, various researchers have demonstrated the versatility and effectiveness of this technology in different domains. Lindemann et al. [25] conducted a comparative analysis of LSTM networks for anomaly detection, evaluating their performance against other machine learning and deep learning models. Musleh et al. [26] innovatively combined LSTM with Stacked Autoencoders to develop a system for Automatic Generation Control in power grids, showcasing the adaptability of LSTM in energy sector applications.

Further expanding the scope of LSTM's applications, Mushtaq et al. [27] introduced a hybrid Intrusion Detection System (IDS) model that synergizes Auto-Encoders with LSTM, achieving a remarkable 89% accuracy in classifying cyber-attacks. This model underscores the potential of combining LSTM with other neural network architectures for enhanced cybersecurity measures.

In the realm of the Internet of Things (IoT), Mahmoud et al. [28] developed an Auto-Encoder LSTM (AE-LSTM) model tailored for detecting anomalies. Notably, their model achieved high levels of accuracy without the need for extensive data pre-processing, indicating the efficiency and robustness of AE-LSTM in handling IoT-related data.

Altunay and Albayrak [20] also proposed a hybrid model integrating CNN (Convolutional Neural Networks) with LSTM for IDS. This model demonstrated significant accuracy in binary and multi-class classifications, highlighting the effectiveness of combining CNN's spatial analysis capabilities with LSTM's temporal data processing strength in creating advanced IDS solutions. These diverse applications illustrate the wide-ranging potential and adaptability of LSTM networks in addressing complex challenges across various technological fields.

Khatan [29] introduced a hybrid DL model combining CNN, LSTM, and a self-attention algorithm, significantly enhancing predictive capabilities across various datasets.

3. Methodology

A deep learning (DL) model incorporating transfer learning was developed, and a pre-trained model was utilized to optimize it for a specific dataset. This approach reduces training time and enhances efficiency. In the initial stages of the neural network, low-level and task-specific features are extracted. The model employs transfer learning in its early and central layers while retraining later ones. Equipped initially with 2D filters, the model integrates a 3D CNN layer to transform these filters into 3D, enabling the processing of frames in 3D blocks. This addition allows for capturing and analyzing temporal and spatial information, particularly useful in video motion analysis.

In contrast to feed-forward neural networks, where each input is processed independently, Recurrent Neural Networks (RNNs) establish a temporal dependency, basing each prediction at time 't' on previous predictions and accumulated knowledge. However, RNNs may fall short in addressing long-term dependencies, a gap effectively filled by Long Short-Term Memory (LSTM) networks. LSTMs utilize cell states to carry information through the network,

selectively retaining or discarding it as needed.

The development of the physical abuse detection model proceeded in three stages:

Stage-1: Pre-trained 2D models, specifically ResNet-18 and GoogleNet, were explored. These models were trained and tested on two datasets using transfer learning, focusing on 2-dimensional filters for spatial information processing.

Stage 2: Incorporation of a 3D CNN layer into the existing 2D models, effectively upgrading the kernel dimensions from 2D to 3D. This enhancement allows for extracting temporal information from the first layer, facilitating efficient spatiotemporal information processing. The modified models, named '3D ResNet-18' and '3D GoogleNet', were then considered.

Stage 3: A bilinear LSTM layer was added to the 3D CNN integrated models of 3D ResNet-18 and 3D GoogleNet to boost the temporal information processing capabilities further. The resulting models are '3D ResNet-18 + LSTM' and '3D GoogleNet + LSTM'.

The framework for detecting physical abuse through these hybrid DL models is outlined in Figure 1. Initially, raw video inputs undergo preprocessing steps like resizing, normalization, and grayscale conversion. Data augmentation techniques are applied to address any skewness in the dataset. The dataset is then split into training (70%), validation (10%), and testing (20%) segments. During training, the model is fine-tuned through hyperparameter adjustments. Training accuracy and loss are monitored through graphs, with learning rate, and epoch numbers tweaked to guide the model toward optimal loss minimization and accuracy maximization, ensuring stability.

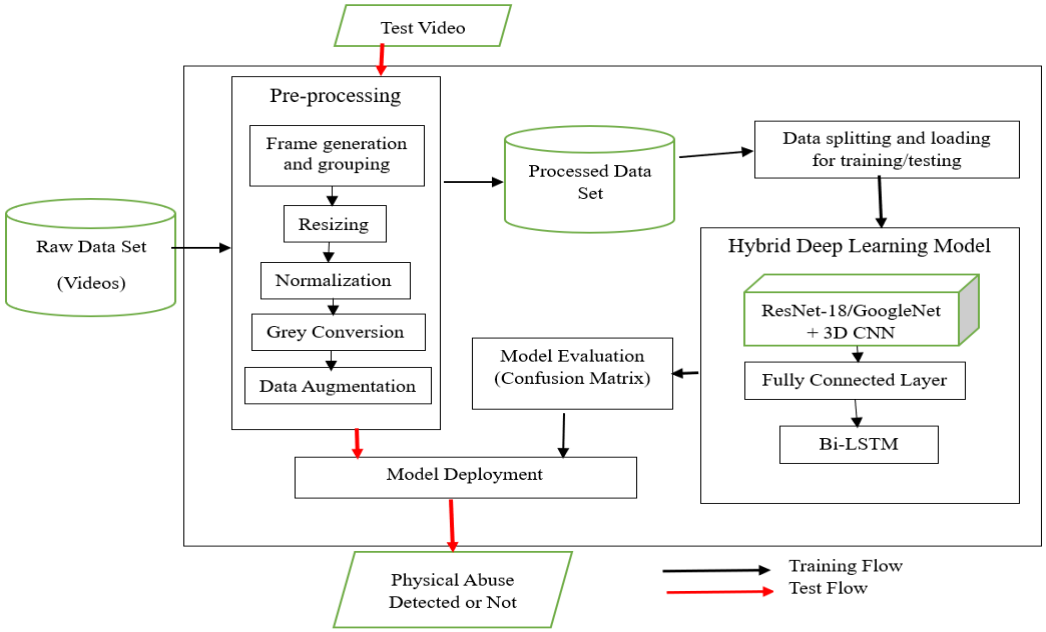

Fig 1: Schematic Representation of the Hybrid Model for Physical Abuse Detection**.**

4. Dataset

Deep Learning (DL) models advance their knowledge by analyzing training datasets, especially in computer vision systems where algorithms are tailored to identify specific features within these datasets. However, sourcing an appropriate dataset for training poses a significant challenge in developing a DL model. The complexity of human interactions, with their varied body movements and coordination, adds to this challenge. An analysis of this variability, particularly in the context of single images, has been conducted by Ronchi and Perona [36].

Human interactions are inherently dynamic, often characterized by coordinated movements. For instance, a simple handshake involves one person extending their hand, followed by the reciprocal action of the other person. Researchers have recognized the significance of understanding and predicting future actions in such scenarios, as it significantly contributes to the overall comprehension of the scene [30]. Some studies have discovered that forecasting future actions can be accomplished by classifying an action or interaction based on its initial stages [31]. While this approach holds promise for scenarios with clear objectives, its effectiveness diminishes as the complexity and variability of interactions increase, especially in more nuanced and socially oriented interactions, such as playful gestures.

In the context of our project, one of the significant challenges we encountered was the acquisition of a suitable dataset directly related to physical abuse. While datasets like UT-Interaction, UCF101, and the Kinetics action recognition dataset, which encompass actions like boxing and punching, exist, they often fall short of meeting the specific requirements for capturing instances of physical abuse in video feeds. These datasets typically offer lower-resolution inputs and may not align with the precise criteria for detecting physical abuse actions in a video context. After conducting an exhaustive search [33], [34], we eventually identified two publicly available datasets, namely UBI-Fights [35] and UCF Crime [36], which contain authentic video clips depicting everyday individuals engaging in actions such as hitting, kicking, and various forms of physical conflict. These datasets closely matched our project's objectives, providing a more appropriate

foundation for detecting physical abuse in video data.

4.1. UBI Fight Dataset

Released in 2020, this dataset encompasses 80 hours of video material, captured using a stationary camera and meticulously annotated at the frame level. It comprises 1,000 video clips, of which 216 depict instances of physical altercations, and 784 portray everyday life scenarios. Augmentation techniques enhanced the dataset, expanding the collection to 2,800 video segments. Each clip lasts between 80 to 90 seconds and is recorded at 30 frames per second. Video frames are resized, normalized, and transformed into grayscale for uniformity and quality in data processing. Illustrations of select video clips and action sequences from this dataset are presented in Figure 2 and Figure 3. The dataset is methodically divided for different phases of the model development: 70% is dedicated to training, 10% to validation, and the remaining 20% is reserved for testing. This partitioning ensures that the dataset is well-organized and representative, facilitating the development and thorough evaluation of a robust model.

Fig 2: Examples of Video Clips from the UBI Fight Dataset.

Fig 3: Image Sequence from a Video Clip in the UBI Fight Dataset.

4.2. UCF Crime Dataset

In 2021, the University of Central Florida introduced a dataset featuring unedited surveillance videos encompassing various unusual events. This extensive dataset includes 128 hours of video footage, which has been augmented to encompass 3,500 video clips. Each clip is roughly 1.5 minutes long, providing a rich resource for research and analytical purposes.

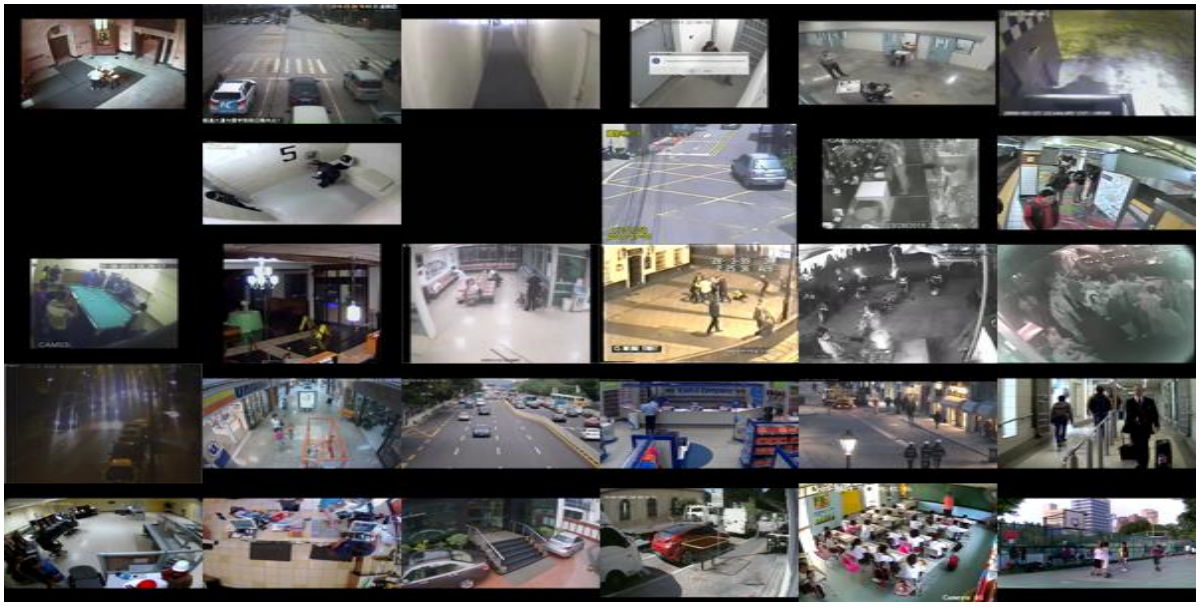

Fig 4: Representative Video Clips from the UCF Crime Dataset.

5. Performance Metric

To evaluate the performance of the Deep Learning (DL) model in detecting physical abuse, a confusion matrix was created using 20% of the dataset reserved for testing. The availability of an annotated dataset enabled the generation of this matrix. The matrix categorizes results into four distinct labels based on comparing actual and predicted values. It serves as a tool to compute key performance metrics such as accuracy, sensitivity, precision, and specificity.

The confusion matrix acts as a summary chart, illustrating how the model's predictions align with the observed outcomes. In this context, accuracy refers to the model's ability to correctly identify physical abuse and recognize when it is not occurring. This means the model's predictions should consistently match physical abuse's actual occurrence or absence for a high accuracy score.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{ .....(1)}$$

Recall, also known as Sensitivity, measures the proportion of actual positives (true positives) correctly identified by the model compared to the total number of positives present in the ground truth. The classifier yields many false negatives if the recall score is less than 0.5. This scenario might arise due to an imbalance in the analyzed classes or the model's hyperparameters not optimally tuned.

$$Recall/Sensitivity = \frac{TP}{TP+FN} \text{ .....(2)}$$

Specificity refers to the ratio of true negatives, which measures how accurately the actual negative cases are identified and predicted as negative by the model.

$$Specificity = \frac{TN}{TN+FP} \text{.....(3)}$$

Precision is the metric that quantifies the ratio of correctly predicted positive observations (true positives) to the total predicted positives. If the precision score falls below 0.5, it suggests that the classifier is generating a significant number of false positives. This issue could stem from an imbalance in the dataset or from hyperparameters of the model that require fine-tuning.

$$Precision = \frac{TP}{TP+FP} \text{.....(4)}$$

The Precision-Recall trade-off refers to the balancing act between improving precision and improving recall in a model. Enhancing one often leads to a decrease in the other, so it's about finding an optimal balance based on the specific requirements and priorities of the model.

6. Observation And Result Analysis

Confusion matrices were created for six models using 20% of UBI Fight and UCF Crime datasets. The UBI Fight dataset, after augmentation, consists of 2,780 video clips. According to the allocation strategy, 70% of these clips are used for training, 10% for validation, and 20% for testing. This means 556 clips (20% of 2,780) are designated for testing. Similarly, the UCF Crime dataset was expanded to 3,515 clips. Applying the same distribution, 703 clips (20% of 3,515) are set aside for testing. Figures 5 and 6 showcase the confusion matrices for each model, with Figure 5 focusing on the UBI Fight dataset and Figure 6 highlighting the UCF Crime dataset, providing a clear visual representation of each model's performance.

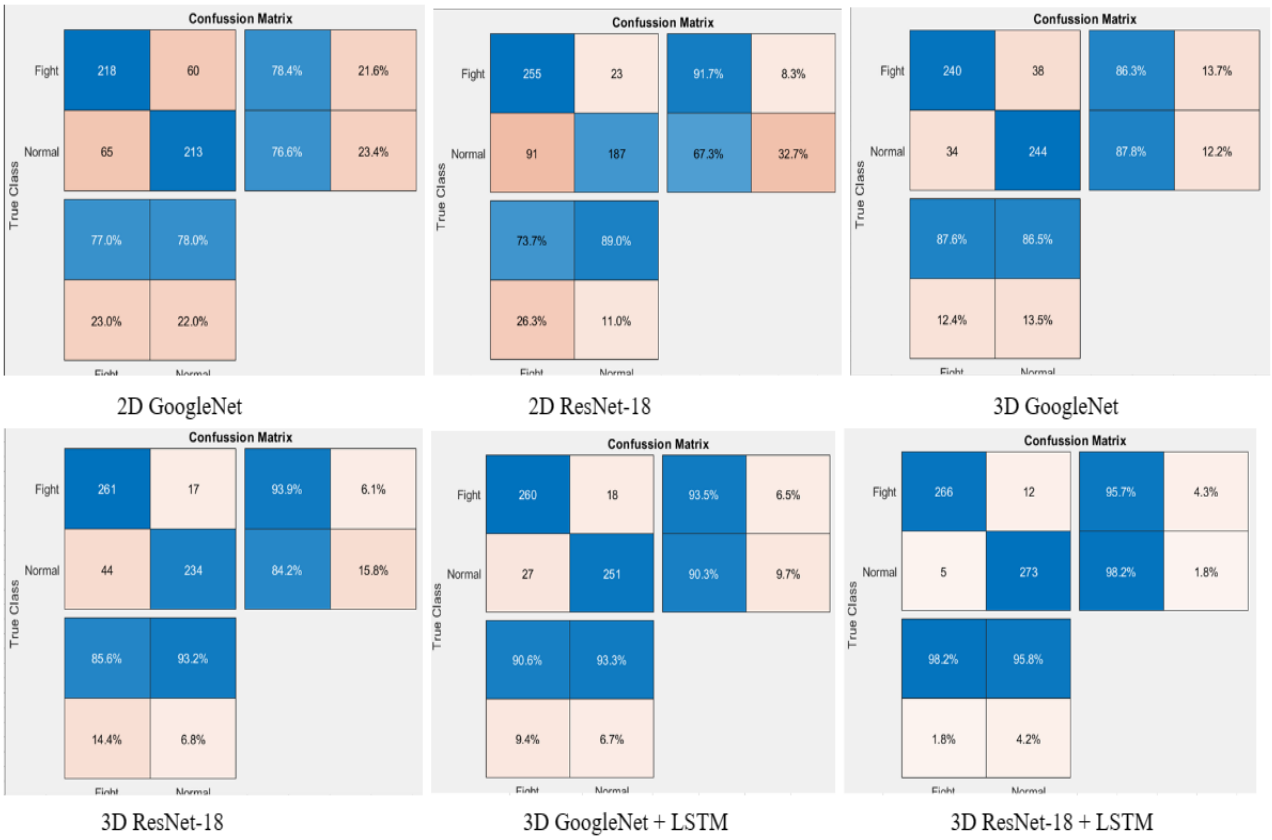

Fig 5: Confusion Matrices for Various Models Analyzed with the UBI Fight Dataset.

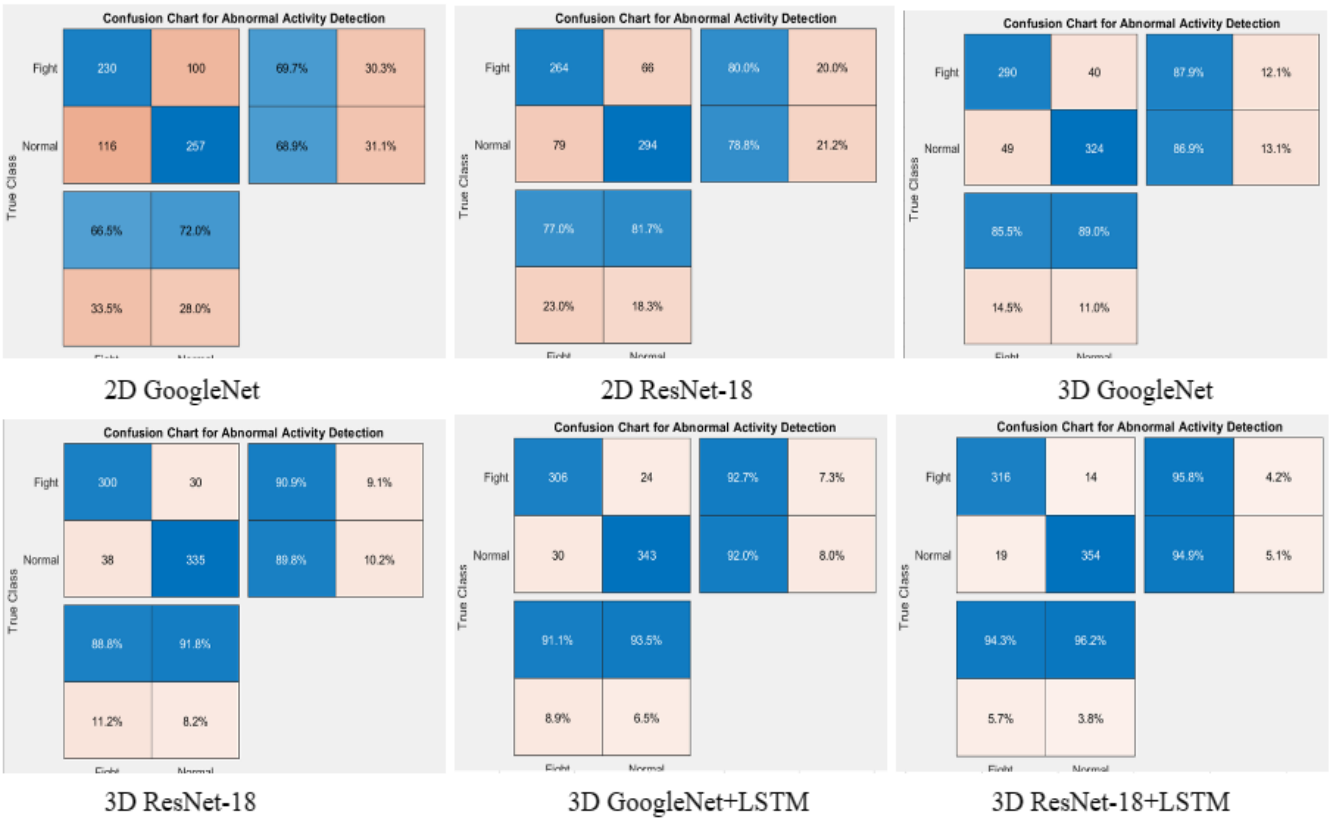

Fig 6: Confusion Matrices for Different Models Evaluated Using the UCF Crime Dataset.

Table 1: Comprehensive Performance Metrics for Six Models Across Two Datasets.

	UBI-Fight				UCF-Crime			
	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision
2D GoogleNet	77.51	78.41	76.61	78.02	69.27	69.69	68.91	71.98
2D ResNet-18	79.49	91.72	67.26	89.04	79.37	80	78.82	81.67
3D GoogleNet	87.05	86.33	87.76	86.52	87.34	87.87	86.86	89.01
3D ResNet-18	89.02	93.88	84.17	93.22	90.32	90.9	89.81	91.78
3D GoogleNet + LSTM	91.9	93.52	90.28	93.3	92.32	92.72	91.95	93.46
3D ResNet-18 + LSTM	96.94	95.68	98.2	95.78	95.31	95.76	94.91	96.2

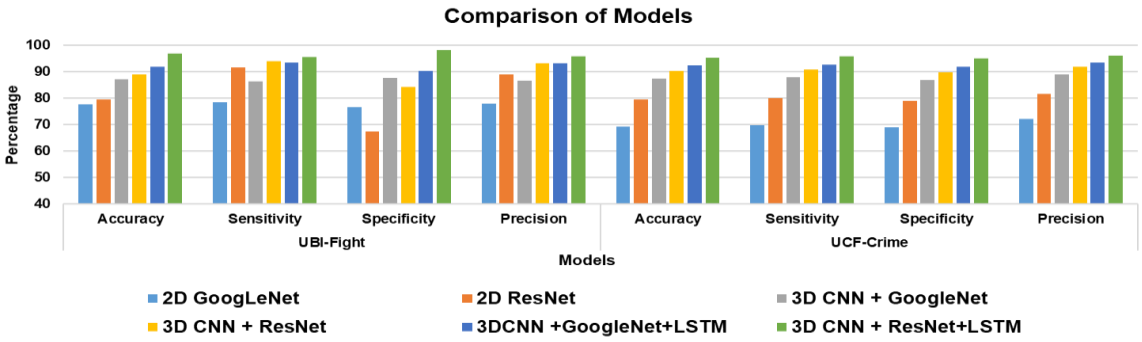

Fig 7: Combined Graphical Display of Performance Metrics Across Six Models and Two Datasets.

Leveraging the information extracted from the confusion matrices, we computed critical performance metrics that are pivotal indicators of the models' effectiveness. These key metrics encompass Accuracy, Precision, Recall, and the F1 Score, each playing a distinct role in assessing the model's performance. The mathematical formulations for these metrics are thoughtfully detailed in equations (1), (2), (3), and (4), all of which can be found in section 5 of the report.

To present a comprehensive overview of the models' performance across different aspects, we have organized the results into a neatly compiled summary table, conveniently labeled Table 1. This table is a centralized reference point, allowing stakeholders to swiftly gauge and compare the models' performance based on the metrics that matter most.

Furthermore, recognizing the significance of visualizing performance trends, we have thoughtfully crafted Figure 7. This graphical representation visually analyzes how the models fare regarding their effectiveness. It offers a clear and intuitive way to discern the strengths and weaknesses of each model, making it easier for decision-makers to draw insights and make informed choices based on the presented data.

6.1. Comparative Analysis of ResNet-based models.

Figures 8 and 9 display the performance metrics' graphical representation across the three implementation stages for the UBI Fight and UCF Crime datasets, respectively. These graphs illustrate a clear trend of improvement in every performance metric as the implementation progresses from stage 1 to stage 3. Notably, the 3D ResNet + LSTM model achieves the highest accuracy, reaching 96.94% for the UBI Fight dataset and 94.91% for the UCF Crime dataset, as depicted in Figures 8 and 9.

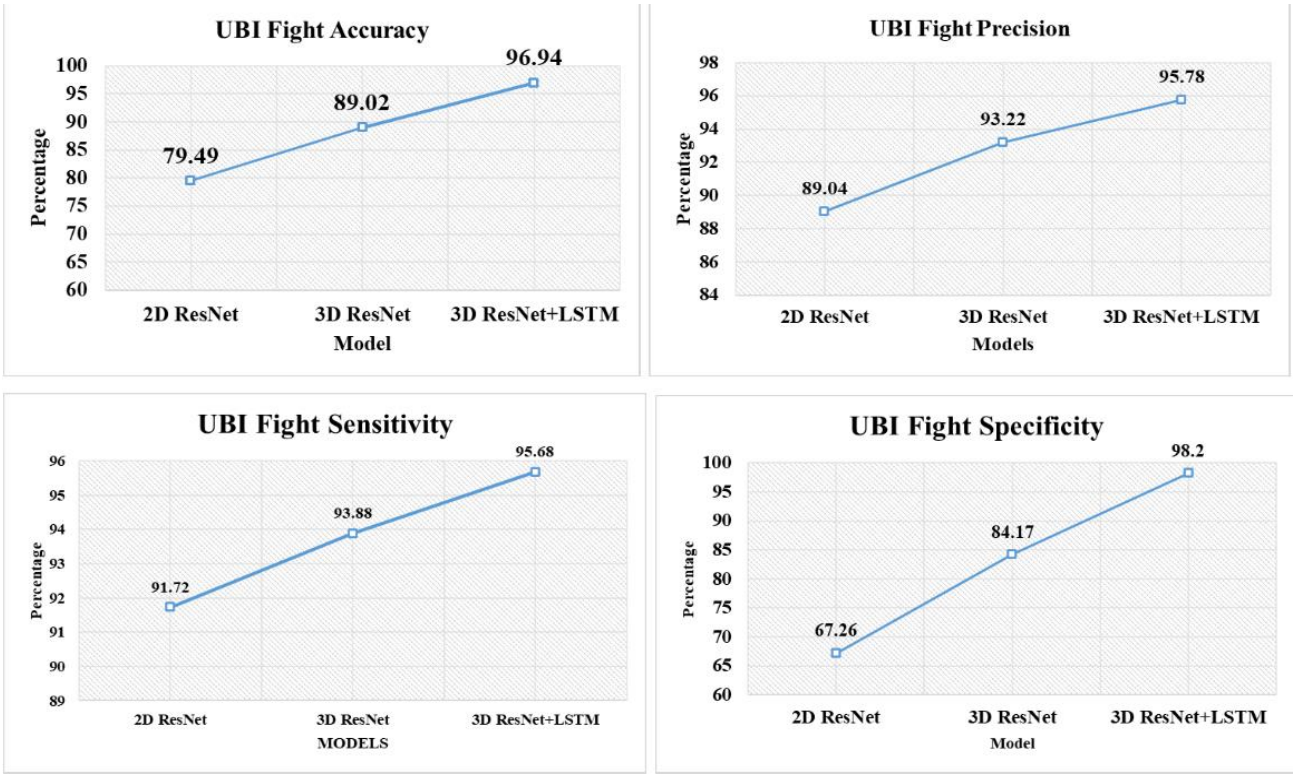

Fig 8: Performance Metrics Visualization for the UBI Fight Dataset Using Three Variants of the ResNet-18 Model.

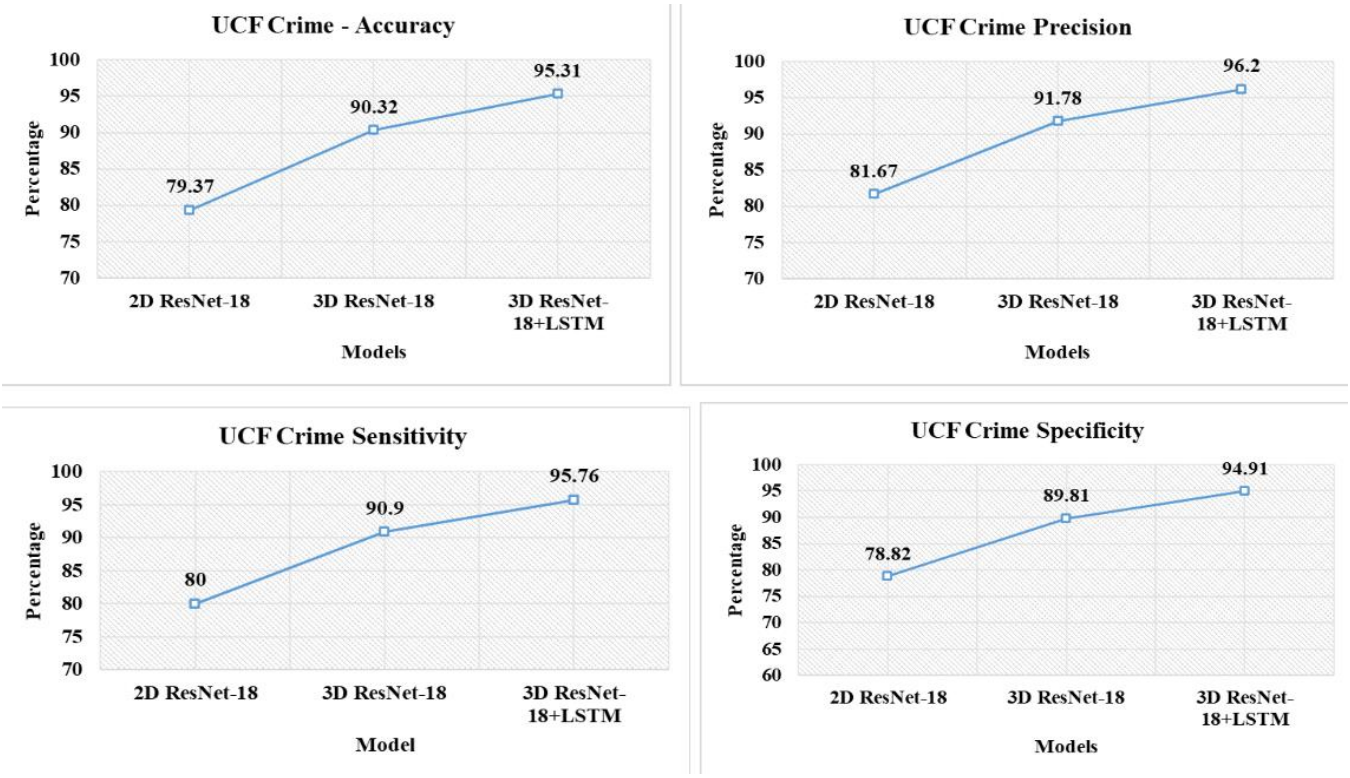

Fig 9: Performance Metrics Graph for the UCF Crime Dataset Using Three Variations of the ResNet-18 Model.

6.2. Comparative Analysis of GoogleNet-based models.

The graphical representation, as illustrated in Figures 10 and 11, demonstrates that for both datasets - UBI Fight and UCF Crime - there is a consistent improvement across all four performance parameters (Accuracy, Precision, Recall, and F1 Score) through the three stages of enhancement applied to the GoogleNet model.

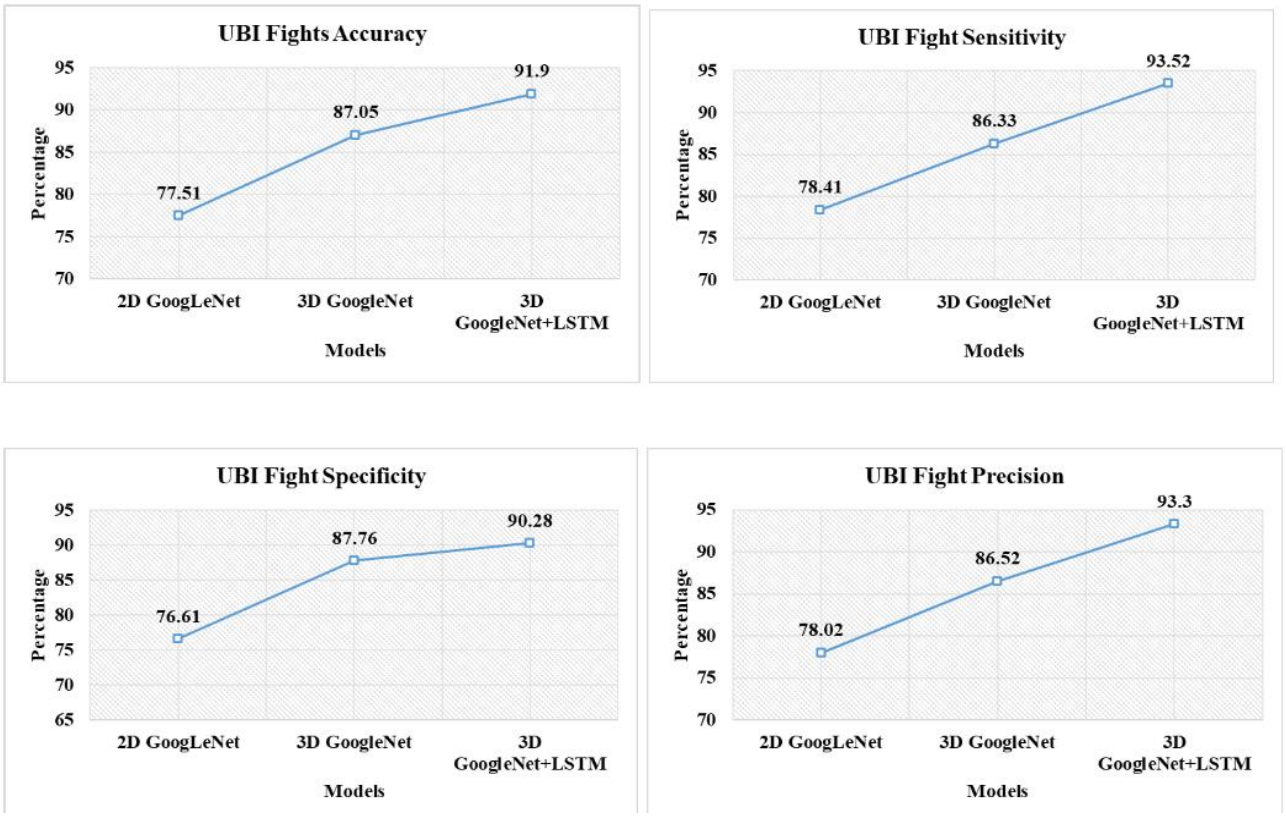

Fig 10: Performance Metrics Visualization for the UBI Fight Dataset Using Three Variants of the GoogleNet Model.

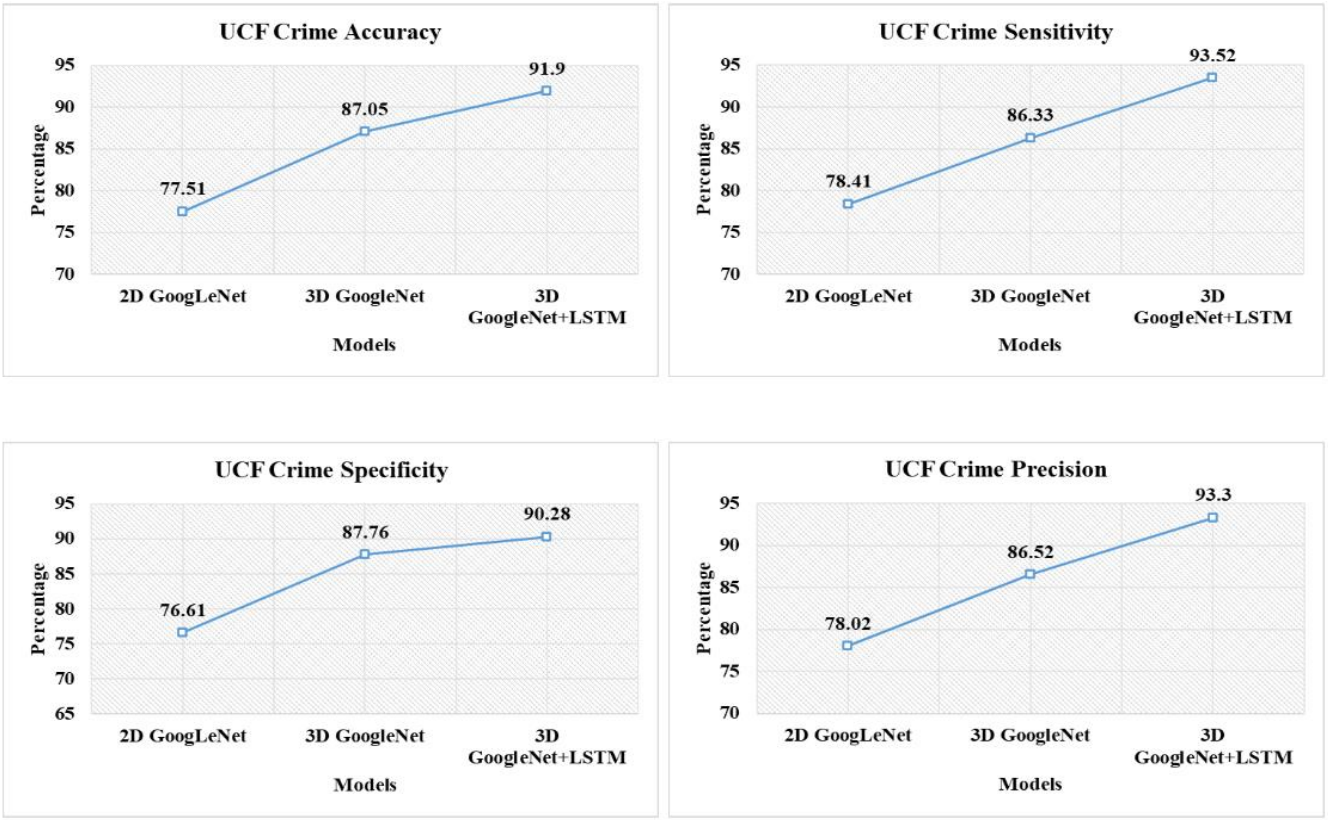

Fig 11: Performance Metrics Graph for the UCF Crime Dataset Using Three Variations of the GoogleNet Model.

6.3. Comparative Analysis of all models

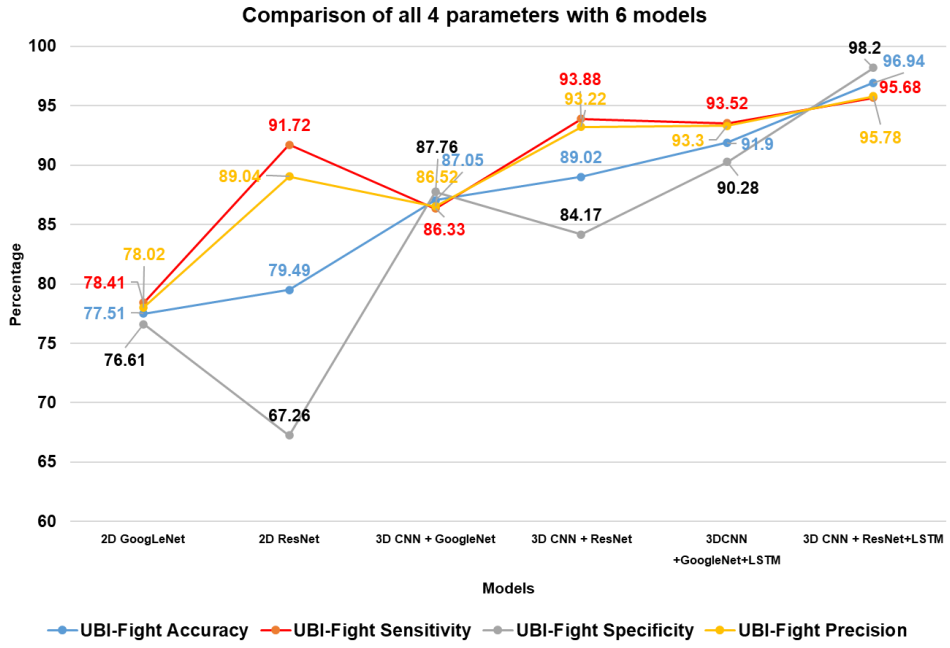

Fig 12: Visualization of Performance Metrics Across All Models.

Figure 12 presents a graphical comparison of the four performance parameters (Accuracy, Precision, Specificity, and Sensitivity) across six models, depicted as four distinct curves. The graph highlights that the 2D ResNet-18 model exhibits more significant fluctuations in these parameter values. In contrast, models based on GoogleNet demonstrate relatively stable performance across all four metrics, showing less variation in accuracy, precision, specificity, and sensitivity scores. While GoogleNet models display a balanced performance, the 3D ResNet-18 + LSTM model stands out by surpassing all others in performance, demonstrating stability in its results.

7. Conclusion And Future Work

CNN-based Deep Learning models excel in extracting and analyzing complex features, with transfer learning demonstrating its effectiveness by quickly stabilizing and delivering strong performance for specific datasets. Incorporating a 3D CNN into these models significantly enhances human action recognition capabilities, as the 3D kernels adeptly process both spatial and temporal data, yielding more accurate results. For intricate tasks like detecting physical abuse, achieving high accuracy, excellent precision, recall, and specificity is crucial.

Adding a bidirectional LSTM layer to the 3D ResNet-18 and 3D GoogLeNet models has further refined the sequential data analysis. The hybrid 3D ResNet-18 with LSTM stands out among the evaluated models, offering superior results while maintaining efficiency.

Looking ahead, there are plans to deploy this model onto an embedded system and test its performance with live video streams in real-time scenarios. The focus will be optimizing throughput and response time to enhance system performance.

Conflicts of interest

The authors of this manuscript state that they have no financial, personal, or professional conflicts of interest that could have influenced the work reported in this paper. This declaration encompasses all forms of potential disputes, ensuring the integrity and impartiality of the research and its findings.

References

[1] Y. Cao et al., "Recognize Human Activities from Partially Observed Videos," in 2013 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Jun. 2013, pp. 2658–2665. doi: 10.1109/CVPR.2013.343.

[2] C. Nolker and H. Ritter, "Visual recognition of continuous hand postures," IEEE Trans Neural Netw, vol. 13, no. 4, pp. 983–994, Jul. 2002, doi: 10.1109/TNN.2002.1021898.

[3] E. Ueda, Y. Matsumoto, M. Imai, and T. Ogasawara, "A hand-pose estimation for vision-based human interfaces," IEEE Transactions on Industrial Electronics, vol. 50, no. 4, pp. 676–684, Aug. 2003, doi: 10.1109/TIE.2003.814758.

[4] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), vol. 37, no. 3, pp. 311–324, May 2007, doi: 10.1109/TSMCC.2007.893280.

[5] A. Mumtaz, A. B. Sargano, and Z. Habib, "Robust learning for real-world anomalies in surveillance videos," Multimed Tools Appl, vol. 82, no. 13, pp. 20303–20322, May 2023, doi: 10.1007/s11042-023-14425-x.

[6] Yanmin Zhu, Zhibo Yang, and Bo Yuan, "Vision-Based Hand Gesture Recognition," in 2013 International Conference on Service Sciences (ICSS), IEEE, Apr. 2013, pp. 260–265. doi: 10.1109/ICSS.2013.40.

[7] S. Waheed, R. Amin, J. Iqbal, M. Hussain, and M. A. Bashir, “An Automated Human Action Recognition and Classification Framework Using Deep Learning,” in 2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), IEEE, Mar. 2023, pp. 1–5. doi: 10.1109/iCoMET57998.2023.10099190.

[8] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, “Human Action Recognition From Various Data Modalities: A Review,” IEEE Trans Pattern Anal Mach Intell, pp. 1–20, 2022, doi: 10.1109/TPAMI.2022.3183112.

[9] D. Liang and E. Thomaz, “Audio-Based Activities of Daily Living (ADL) Recognition with Large-Scale Acoustic Embeddings from Online Videos,” Proc ACM Interact Mob Wearable Ubiquitous Technol, vol. 3, no. 1, pp. 1–18, Mar. 2019, doi: 10.1145/3314404.

[10] D. Ganesh, R. R. Teja, C. D. Reddy, and D. Swathi, “Human Action Recognition based on Depth maps, Skeleton and Sensor Images using Deep Learning,” in 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), IEEE, Oct. 2022, pp. 1–8. doi: 10.1109/GCAT55367.2022.9971982.

[11] M. DALLEL, V. HAVARD, D. BAUDRY, and X. SAVATIER, “InHARD - Industrial Human Action Recognition Dataset in the Context of Industrial Collaborative Robotics,” in 2020 IEEE International Conference on Human-Machine Systems (ICHMS), IEEE, Sep. 2020, pp. 1–6. doi: 10.1109/ICHMS49158.2020.9209531.

[12] J.-S. Kim, “Efficient Human Action Recognition with Dual-Action Neural Networks for Virtual Sports Training,” in 2022 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), IEEE, Oct. 2022, pp. 1–3. doi: 10.1109/ICCE-Asia57006.2022.9954758.

[13] P. Le Noury, R. Polman, M. Maloney, and A. Gorman, “A Narrative Review of the Current State of Extended Reality Technology and How it can be Utilised in Sport,” Sports Medicine, vol. 52, no. 7, pp. 1473–1489, Jul. 2022, doi: 10.1007/s40279-022-01669-0.

[14] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, “A new hybrid deep learning model for human action recognition,” Journal of King Saud University - Computer and Information Sciences, vol. 32, no. 4, pp. 447–453, May 2020, doi: 10.1016/j.jksuci.2019.09.004.

[15] H.-H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, “Exploiting deep residual networks for human action recognition from skeletal data,” Computer Vision and Image Understanding, vol. 170, pp. 51–66, May 2018, doi: 10.1016/j.cviu.2018.03.003.

[16] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jul. 2017, pp. 1302–1310. doi: 10.1109/CVPR.2017.143.

[17] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, “Two-person interaction detection using body-pose features and multiple instance learning,” in 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, Jun. 2012, pp. 28–35. doi: 10.1109/CVPRW.2012.6239234.

[18] R. A. Guler, N. Neverova, and I. Kokkinos, “DensePose: Dense Human Pose Estimation in the Wild,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Jun. 2018, pp. 7297–7306. doi: 10.1109/CVPR.2018.00762.

[19] Ho Yub Jung, Soochahn Lee, Yong Seok Heo, and Il Dong Yun, “Random tree walk toward instantaneous 3D human pose estimation,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2015, pp. 2467–2474. doi: 10.1109/CVPR.2015.7298861.

[20] H. C. Altunay and Z. Albayrak, “A hybrid CNN+LSTM-based intrusion detection system for industrial IoT networks,” Engineering Science and Technology, an International Journal, vol. 38, p. 101322, Feb. 2023, doi: 10.1016/j.jestch.2022.101322.

[21] A. Raza et al., “A Hybrid Deep Learning-Based Approach for Brain Tumor Classification,” Electronics (Basel), vol. 11, no. 7, p. 1146, Apr. 2022, doi: 10.3390/electronics11071146.

[22] R. Tandon, S. Agrawal, A. Chang, and S. S. Band, “VCNet: Hybrid Deep Learning Model for Detection and Classification of Lung Carcinoma Using Chest Radiographs,” Front Public Health, vol. 10, Jun. 2022, doi: 10.3389/fpubh.2022.894920.

[23] V. Hnamte, H. Nhung-Nguyen, J. Hussain, and Y. Hwa-Kim, “A Novel Two-Stage Deep Learning Model for Network Intrusion Detection: LSTM-AE,” IEEE Access, vol. 11, pp. 37131–37148, 2023, doi: 10.1109/ACCESS.2023.3266979.

[24] Mst. A. Khatun et al., “Deep CNN-LSTM With Self-Attention Model for Human Activity Recognition Using Wearable Sensor,” IEEE J Transl Eng Health Med, vol. 10, pp. 1–16, 2022, doi: 10.1109/JTEHM.2022.3177710.

[25] B. Lindemann, B. Maschler, N. Sahlab, and M. Weyrich, “A survey on anomaly detection for technical systems using LSTM networks,” Comput Ind, vol. 131, p. 103498, Oct. 2021, doi: 10.1016/j.compind.2021.103498.

[26] A. S. Musleh, G. Chen, Z. Y. Dong, C. Wang, and S. Chen, “Attack Detection in Automatic Generation Control Systems using LSTM-Based Stacked Autoencoders,” IEEE Trans Industr Inform, vol. 19, no. 1, pp. 153–165, Jan. 2023, doi: 10.1109/TII.2022.3178418.

[27] E. Mushtaq, A. Zameer, M. Umer, and A. A. Abbasi, “A two-stage intrusion detection system with auto-encoder and LSTMs,” Appl Soft Comput, vol. 121, p. 108768, May 2022, doi: 10.1016/j.asoc.2022.108768.

[28] M. Mahmoud, M. Kasem, A. Abdallah, and H. S. Kang, “AE-LSTM: Autoencoder with LSTM-Based

Intrusion Detection in IoT," in 2022 International Telecommunications Conference (ITC-Egypt), IEEE, Jul. 2022, pp. 1–6. doi: 10.1109/ITC-Egypt55520.2022.9855688.

[29] Mst. A. Khatun et al., "Deep CNN-LSTM With Self-Attention Model for Human Activity Recognition Using Wearable Sensor," IEEE J Transl Eng Health Med, vol. 10, pp. 1–16, 2022, doi: 10.1109/JTEHM.2022.3177710.

[30] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating Visual Representations from Unlabeled Video," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2016, pp. 98–106. doi: 10.1109/CVPR.2016.18.

[31] M. Ziaeefard, R. Bergevin, and L.-P. Morency, "Time-slice Prediction of Dyadic Human Activities," in Proceedings of the British Machine Vision Conference 2015, British Machine Vision Association, 2015, pp. 167.1-167.13. doi: 10.5244/C.29.167.

[32] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in 2011 International Conference on Computer Vision, IEEE, Nov. 2011, pp. 1036–1043. doi: 10.1109/ICCV.2011.6126349.

[33] F. J. Rendón-Segador, J. A. Álvarez-García, J. L. Salazar-González, and T. Tommasi, "CrimeNet: Neural Structured Learning using Vision Transformer for violence detection," Neural Networks, vol. 161, pp. 318–329, Apr. 2023, doi: 10.1016/j.neunet.2023.01.048.

[34] M. A. B. Abbass and H.-S. Kang, "Violence Detection Enhancement by Involving Convolutional Block Attention Modules Into Various Deep Learning Architectures: Comprehensive Case Study for UBI-Fights Dataset," IEEE Access, vol. 11, pp. 37096–37107, 2023, doi: 10.1109/ACCESS.2023.3267409.

[35] C. Leng, Q. Ding, C. Wu, and A. Chen, "Augmented two-stream network for robust action recognition adaptive to various action videos," J Vis Commun Image Represent, vol. 81, p. 103344, Nov. 2021, doi: 10.1016/j.jvcir.2021.103344.

[36] M. Z. Zaheer, A. Mahmood, H. Shin, and S.-I. Lee, "A Self-Reasoning Framework for Anomaly Detection Using Video-Level Labels," IEEE Signal Process Lett, vol. 27, pp. 1705–1709, 2020, doi: 10.1109/LSP.2020.3025688.

Bibliography

Prof. Srividya M. S., an Assistant Professor in the Computer Science and Engineering Department at R.V College of Engineering, boasts a distinguished career with over 12 years of teaching experience and an 8-year tenure in the industry. Her primary research interests lie in Deep Learning, Artificial Intelligence, Machine Learning, and Computer Vision. She has mentored numerous undergraduate projects and contributed extensively to international journals and conferences. Additionally, she has been involved in various research and consultancy projects.

Dr. Anala M. R., a Professor in the Information Science and Engineering Department at R.V College of Engineering, brings over 22 years of teaching experience to her role. Her research primarily focuses on computer architecture, high-performance computing, distributed systems, parallel programming, computer vision, and deep learning. With an impressive record, she has supervised more than 45 undergraduate and 25 postgraduate projects. Her contributions extend to numerous publications in international journals and conferences. She leads the Visual Computing Centre of Excellence and has been actively involved in various research and consultancy projects.