Information Processing and Management 63 (2026) 104439

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Attention-guided residual shrinkage with gated recurrent unit for human activity recognition

S. Banushri [a,b,*], R. Jagadeesha [a,b]

[a] Department of Computer Science and Engineering, Impact College of Engineering and Applied Sciences, Banglore 560092, India
[b] Visvesvaraya Technological University, Belagavi 590018, India

ARTICLE INFO

Keywords:
Gated recurrent unit
Human action recognition
Multiple residual shrinkage building units
Spatial features
Soft threshold and temporal features

ABSTRACT

Human Action Recognition (HAR) is crucial for applications like video surveillance. Although numerous algorithms have been developed for HAR, these algorithms have failed to appropriately extract spatial and temporal features. In this manuscript, a HAR architecture is developed that integrates Multiple Residual Shrinkage Building Units (MRSBU) for spatial feature extraction with Gated Recurrent Unit (GRU) for temporal modeling. The architecture employs Inception v3 to capture rich spatial features from video frames, and the MRSBU learns complex temporal features using an adaptive soft-thresholding mechanism to suppress noisy and redundant features. This is then fed into a GRU with a temporal attention mechanism that applies dynamic significance to each frame and classifies the human activities. This lightweight and effective model demonstrates superior generalization across multiple challenging HAR dataset. The proposed MRSBU with GRU algorithm achieved an accuracy of 99.75 % on the UCF50 dataset, 99.55 % on the UCF101 dataset, and 98.95 % on the HMDB51 dataset, outperforming conventional Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM), Vision Transformer (ViT), and 3D-CNN models. These results show the proposed model's effectiveness and robustness across different video scenarios, including real-world surveillance and anomaly detection applications.

1. Introduction

Human Action Recognition (HAR) has become a crucial domain of research in computer vision, involving applications such as video surveillance (Thakur et al., 2022, Paula et al., 2023) and pedestrian detection (Pareek et al., 2024, Lin & Tseng, 2024, Bilal et al., 2022). It includes classifying human actions in videos into their respective activity classes, which is a complex task because of the spatiotemporal nature of videos (Kushwaha et al., 2023, Khaliluzzaman et al., 2024). The challenges include occlusion, similar actions, and varying durations of human activities (R. Vrskova et al., 2022). While analyzing a video, human actions are differentiated by movements of different body parts (Ahmed et al., 2024). Unlike static images (Khan et al., 2025), individual frame cannot fully represent an action (Hassan et al., 2024, Dharejo et al., 2022). For instance, the pose of a football player (Aggarwal et al., 2024) and jumping for a header resemble that of a person skipping rope (Kumar & Kumar, 2023), only the frame sequence reveals a true activity.

In non-stationary data streams, the performance of previously trained methods tends to decline when encountering changes or differences in newly arriving data, as they struggle to adapt to evolving patterns (Pandey & Kumar, 2024). Adaptability toward new data distributions is essential to address these challenges, as it requires diversity in non-stationary environments (Karim et al., 2024).

* Corresponding author.
E-mail addresses: Banushri914@gmail.com (S. Banushri), jagdish.mtech@gmail.com (R. Jagadeesha).

https://doi.org/10.1016/j.ipm.2025.104439
Received 13 June 2025; Received in revised form 6 October 2025; Accepted 6 October 2025
Available online 13 October 2025

Numerous previous approaches for HAR were based on human silhouettes. The outline of a person across a sequence of frames was captured by evaluating, decomposing (Liu et al., 2024), and subtracting the background noise to discriminate among various activities (Yosry et al., 2024). Temporal data were acquired by tracking movements of human body parts through the integration of captured silhouettes (Jlidi et al., 2025, R Vrskova et al., 2022). Although these algorithms capture the local content of human activities, they are unsuitable in situations where multiple persons perform certain activities simultaneously (Bharathi & Sridevi, 2023).

Instead of conventional algorithms, several Deep Learning (DL)-enabled methods (Javed et al., 2022) have been developed for HAR (Dastbaravardeh et al., 2024). The primary aim of computer vision researchers in DL-based HAR is to deliver improved performance through DL in different domains, such as face recognition (Zakariah & Alnuaim, 2024), image recognition (Amrani et al., 2025), person re-identification (Rani & Kumar, 2024), and object detection (Bhavani & Pagi, 2025, Gangwar et al., 2024, Ullah et al., 2023). Fig. 1 represents the system architecture of HAR.

1.1. Problem statement

Existing HAR models face challenges due to feature redundancy, noisy activations, and temporal misalignments. Traditional convolutional architectures use irrelevant or low-confidence spatial features in deeper layers, reducing discriminative accuracy in scenes with occlusions and multiple overlapping actions. Moreover, traditional recurrent models suffer from vanishing gradients and lack precision in capturing long-range dependencies when action transitions are subtle. These challenges result in suboptimal recognition accuracy, poor generalization across dataset, and reduced robustness in unconstrained environments.

1.2. Research objective

The primary objective of this article is to develop a robust and effective HAR method that addresses challenges such as noisy features, redundant activations, and temporal misalignments. To achieve this, the proposed model incorporates a Multiple Residual Shrinkage Building Units (MRSBU) module that suppresses noisy and redundant activations through a soft-thresholding mechanism, improving feature relevance and spatial discrimination. The Gated Recurrent Unit (GRU) component is designed to capture long-range temporal dependencies with minimized computational cost. Moreover, the method is intended to generalize well to real-world scenarios, as demonstrated by its evaluation on standard and challenging video dataset such as UCF50, UCF101, HMDB51, UCF Crime, and UCSD.

1.3. Contribution

The significant contributions of this manuscript are described below.

MRSBU with GRU algorithm is developed for efficient implementation of HAR.

Attention-guided residual shrinkage blocks are incorporated to suppress noise using a soft-thresholding mechanism. While standard residual blocks refine features, RSBU learns a dynamic threshold that determines which features are significant for HAR.

The output of RSBU is flattened and arranged across time steps, and these sequences are fed into the GRU layer to learn temporal relationships. The GRU helps capture the transitions and dynamics of movement patterns.

The rest of the sections in this article are organized as follows: Section 2 analyzes existing approaches with their advantages and limitations. Section 3 explains the proposed approach. Section 4 discusses the results and comparisons of the proposed and previously developed approaches. Finally, Section 5 concludes the article.

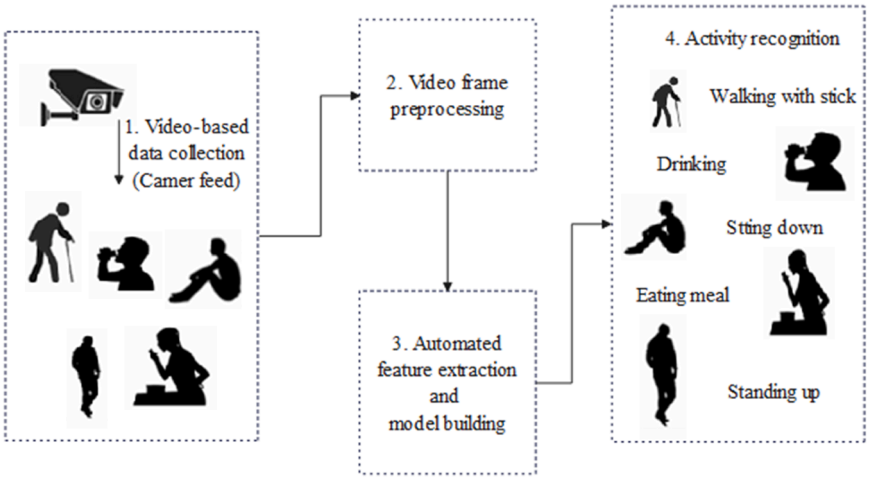

Fig. 1. Workflow of human activity recognition system from video capture to activity classification.

2. Literature review

Activity recognition tasks involve detecting activities from videos that contain action sequences over a duration. Unlike image processing, video processing requires both spatiotemporal and spatial data.

2.1. CNN-based models

Varshney and Bakariya (Varshney & Bakariya, 2022) suggested a Deep Convolutional Neural method to recognize human activities in videos by integrating multiple CNN streams. The suggested method combined spatial and temporal data through double fusion mechanisms, namely average fusion and convolution fusion of spatial and temporal features. Khan et al. (Khan et al., 2024). presented a fully automated scheme for HAR through integration of a Deep Neural Network (DNN) and multiple view features. DNN features were extracted using a pre-trained CNN model, VGG19. Multiple view features were computed from horizontal and vertical gradients, using vertical direction features. Parameters such as mutual information, relative entropy, and Strong Correlation Coefficient (SCC) were employed to select optimal features using a threshold function. The chosen features were then fed to Naive Bayes classifier for recognition. Gupta and Mahapatra (Gupta & Mahapatra, 2025) introduced the Hybrid Efficient Xception Network (HyEx-Net) for HAR. The input color image was used for feature extraction, while depth images were processed to extract skeleton joints. Captured skeleton joints were given for local feature extraction, wherein features such as displacement vectors and relative positions were captured. Features from the input depth and color images were then fused and processed using HyEx-Net, which integrated Xception and Efficient networks. Zhao et al. (Zhao et al., 2023). suggested a Three-Dimensional Convolutional Neural Network integrating Channel Attention (3DCCA) for HAR. RGB video frames were pre-processed using mean normalization. 3DCNN was developed for spatio-temporal feature extraction, and Channel Attention (CA) was applied to select essential features. A softmax classifier was used for classification and detection of human behavior from the videos. Lim et al. (Lim et al., 2023). developed Fine-tuned Temporal Dense Sampling using a One-Dimensional –Convolutional Neural Network (1D CNN) (FTDS-1DConvNet) for HAR. This model employed temporal segmentation and dense sampling. Inception-ResNet-V2 was used for fine-tuning segments, with max pooling applied along the temporal axis for encoding the most essential features at fixed length representations. Representations were then processed by a 1DConvNet for representation learning and classification.

2.2. RNN/LSTM/GRU based models

Jayamohan and Yuvaraj (Jayamohan & Yuvaraj, 2025) employed multiple feature extraction and optimization algorithms to improve the accuracy and effectiveness of human action detection. Static and dynamic video components were captured, while sequence modeling and classification were handled by a GRU with an attention mechanism. Grasshopper Optimization Algorithm (GOA) was used for optimizing feature selection and classification phases. Kumar et al. (Praveenkumar et al., 2023). developed a methodology for real-time HAR in compressed video domains using motion vectors and attention-guided bidirectional LSTM (MVABLSTM). The videos were in compression formats of MPEG-4 and H.264 in this study. Tan et al. (Tan et al., 2022). presented a Bidirectional LSTM with Temporal Dense Sampling and Fusion Network model for HAR. Videos were segmented, pooled temporally, and processed using a multi-stream BiLSTM to capture forward and backward long-term spatial and temporal dependencies. Fixed weights were applied to spatial and temporal streams, and a fusion network was trained for final classification.

2.3. Transformer-based models

Yadav et al. (Yadav et al., 2024). employed DL-based transformer models such as Vision Transformer (ViT), Data-efficient Image Transformers (DieT), and SwinV2 for image-based dataset, while VideoMAE was used for video-enabled dataset. Wensel et al. (Wensel et al., 2023). introduced two transformer neural networks for HAR: Recurrent Transformer (ReT), for sequential predictions, and ViT for extracting salient attributes from images to enhance speed and scalability during HAR.

2.4. Hybrid models

Hussain et al. (Hussain et al., 2024). developed an AI-based behavior biometrics architecture using a Dynamic Attention Fusion Unit (DAFU) through a Temporal-Spatial Fusion (TSF) scheme for effective recognition of human activity. In the initial stage, the lightweight EfficientNetB0 was improved through DAFU for extracting salient features via a specified channel-spatial attention mechanism. In the next stage, DAFU attributes along fixed sequence lengths were provided to the developed TSF network for capturing temporal, spatial, and behavioral dependencies. Echo-ConvLSTM, which is a combination of an echo state network and ConvLSTM within the TSF framework, improved accuracy and robustness by integrating temporal dependencies from both the echo state network and ConvLSTM. Sharma and Singh (Sharma & Singh, 2023) introduced spatiotemporal LSTM algorithms named Convolutional Spatio-Temporal Long Short-Term Memory Network (ConvST-LSTM-Net) for skeleton-based action recognition. The main aim of the developed method was to identify significant keypoints in every frame. The ConvST-LSTM-Net method delivered better performance compared with previous methods that utilized various modalities. Ahmad et al. (Ahmad et al., 2023). developed a method for HAR to process visual data through CNN and Bidirectional Gated Recurrent Unit (BiGRU). Initially, deep features were extracted from frame sequences of human activity videos using a CNN. Essential features were then selected from the deep features to enhance performance and minimize computational complexity. For learning temporal motions of frame sequences, the developed Bi-GRU processed those

essential features and learned temporal dynamics for both forward and backward directions at every time step. Gocmen and Murateminakata (Göçmen & Akata, 2023) introduced a new image form known as Poly Sillhouette (PoS), a new Polygonization (PoG) technique, and Polygon Coding (PoC) for HAR. The Polygonization phase relied on curve and image polygonization, which provided fast, adaptable, and precise contour coordinates of PoS images. PoCs were produced by projecting each edge vector derived from corner coordinates into angular regions, after which PoS codes were developed. These coders were grouped as k-mers, similar to the approach of Genetic Algorithms (GA), and were then utilized as features. Uddin et al. (Uddin et al., 2024). implemented the CNN, ConvLSTM, and Long-term Recurrent Convolutional Network (LRCN) frameworks. These methods were introduced for capturing spatial features and temporal dependencies in video information, thereby improving the accuracy of activity classification. The below Table 1 represents a comparative analysis of existing HAR methods.

The proposed MRSBU-GRU architecture addresses the challenges across all categories. By combining the residual shrinkage mechanism, it filters out less relevant spatial features before temporal modeling. Unlike traditional or transformer-based models, this model provides better noise suppression, lower computational overhead, and stronger generalization on benchmark and real-world dataset.

2.5. Model selection

The models in this article are selected to efficiently handle both spatial and temporal characteristics of human activity data in videos. Inception v3 is choosen as the base feature extractor due to its ability to learn complex multi-scale visual patterns with minimized computational cost, thereby addressing spatial feature extraction challenges. To reduce noise in deep CNN features, MRSBU is included for refinement through a dynamic soft-threshold mechanism with attention modules.

For modelling temporal dependencies inherent in video sequences, the GRU network is selected over conventional RNNs and LSTMs because of its simpler architecture, faster training convergence, and effective handling of long-range dependencies with minimized computational overhead. GRUs also overcome the vanishing gradient problem common in deep sequential methods. Additionally, an attention mechanism is integrated with GRU to enhance the model's focus on relevant temporal segments and improve the discriminative power of the final output.

Table 1
Comparative analysis of existing HAR methods highlights dataset, strengths, limitations and innovations addressed by proposed MRSBU-GRU.

Category	Authors and Year	Dataset	Methodology	Accuracy (%)	Advantages	Limitations	Gap Addressed by MRSBU-GRU
CNN-based models	Varshney and Bakariya (Varshney & Bakariya, 2022)	Custom Video	Multi-CNN Streams + Fusion	UCF 101 –97.2 % and HMDB 51 –85.1 %	Captures spatial and temporal features through fusion	No temporal models, lacks feature refinement	Introduced the spatial filtering through soft thresholding
	Khan et al. (Khan et al., 2024).	KTH	VGG19 + Gradient-based Feature Fusion	HMDB 51 –93.7 %	Effective multi-view feature integration	High feature redundancy, lacks sequence modeling	Dynamic feature suppression before temporal modeling
	Gupta and Mahapatra (Gupta & Mahapatra, 2025)	NTU RGB+*D*	HyEx-Net (Xception + EfficientNet)	UCF50 –98.98 %, UCF 101 –98.73 % and HMDB 51 –80.34 %	Uses Multimodal input	Not used noise filtering, no time-series context	MRSBU enhances spatial purity, GRU learns time-dependency
RNN/LSTM/ GRU –based models	Jayamohan and Yuvaraj (Jayamohan & Yuvaraj, 2025)	UCF101	GRU + Attention + GOA	UCF 101 –98.35 % and HMDB 51 –83.45 %	Optimized classification	Limited robustness in real-world data	MRSBU-GRU is lightweight, interpretable and robust
	Kumar et al. (Praveenkumar et al., 2023).	MPEG4, H264	Motion Vectors + Bi-LSTM	UCF 101 –99.6 % and HMDB 51 –90.2 %	Process in compressed domain	Limited to particular formats, lacks spatial adaptively	MRSBU-GRU employs to general video formats with good coverage
Transformer-based models	Yadav et al. (Yadav et al., 2024).	UCF 101, HMDB51	ViT, VideoMAE	UCF 101 –90.9 % and HMDB 51 –59.5 %	Strong global temporal modeling	High computational cost, poor real-time feasibility	MRSBU-GRU balances performance with efficiency
	Wensel et al. (Wensel et al., 2023).	UCF101	ViT-ReT	UCF 101 –94.78 % and HMDB 51 –70.72 %	High scalability in large setups	Needs large-scale pre-training	No pre-training required, efficient in low-resource settings
Hybrid models	Hussain et al. (Hussain et al., 2024).	UCF101	EfficientNetB0 + DAFU + EchoConvLSTM	UCF 101 –71.89 % and HMDB 51 –91.79 %	Integrates attention with temporal modeling	Complex architecture, leads to overfitting	MRSBU-GRU is lightweight, interpretable and robust
	Ahmad et al. (Ahmad et al., 2023).	UCF101	CNN + Bi-GRU	UCF 101 –88.43 % and HMDB 51 –56.23 %	Better temporal learning, minimized complexity	Sensitive to noisy CNN features, limited generalization	MRSBU ensures good feature quality for GRU input

The integrated architecture of Inception v3 for feature extraction, MRSBU for spatial feature enhancement, and GRU with attention for temporal modeling is selected based on its superior balance of accuracy, efficiency, and generalization, as demonstrated through experimental evaluations on multiple HAR dataset.

3. Dataset

The UCF50 (UCF101 dataset, 2025), UCF101 (UCF50 dataset, 2025), and HMDB51 (HMDB51 dataset, 2025) dataset is used in this study. These publicly available dataset for HAR is described below with samples.

3.1. UCF101 dataset

The UCF101 dataset is widely used in HAR research. It contains 13,320 videos clips across 101 action classes, grouped into five broad categories: camera motion, scale, pose, viewpoint, and lighting condition. These diverse conditions make it a suitable dataset for efficient model evaluation. Recordings of 101 activities are collected and classified into 25 groups, each containing 4–7 action videos. Fig. 2 represents sample images from the UCF 101 dataset.

3.2. UCF50 dataset

UCF50 is a moderately sized dataset that consists of 6681 video clips categorized into 50 different classes of human actions, with each clip lasting 2–3 s. Each class contains recordings of different human actions captured from several perspectives, varying in background and camera motion. Fig. 3 presents sample images from the UCF101 dataset.

3.3. HMDB51 dataset

The HMDB51 dataset contains 6849 video clips across 51 action classes, collected from multiple sources such as YouTube, movie clips, the Prelinger archives, and Google Videos. The resolution size of all three dataset is 320 $\times$ 240. Figs. 4 and 5 present sample images from HMDB51 and the comparitive characteristics of all three dataset.

Dataset such as UCF50, UCF101, and HMDB51 are significant for multi-class human action recognition, where the model visually distinguishes between visually similar actions in relatively structured environments. These dataset experiment the model's ability to capture fine-grained temporal and spatial variations in human motion. Additionally, dataset such as UCF-Crime and UCSD Ped2 are designed for anomaly detection in real-world, unconstrained surveillance footage, where abnormal behaviors are temporally spaced and embedded in noise. Using these dataset enables comprehensive testing of the model's ability to classify different actions and to identify subtle behavioural deviations under dynamic and noisy conditions. All three dataset features variability in pose, background clutter, occlusion, and camera motion, which challenge model robustness. By using these three dataset, the model is evaluated on both inter-class and intra-class variability. The exclusion of posture normalization and lighting conditions is a concise methodological decision aligned with the principles of the proposed model. Specifically, Inception v3 inherently manages environmental variations through its multi-scale convolutional architecture, which integrates parallel 1×1, 3×3, 5×5 kernels to efficiently capture hierarchical spatial features that are robust to lighting changes. Batch normalization further stabilizes activations across illumination

Fig. 2. Sample video frames from the UCF101 dataset, which contains 101 action classes across wide range of human activities. The images represent examples from five selected classes, such as Archery, Handstand Walking, Hammer Throw, Golf Swing, and Jump Rope.

Fig. 3. Sample video frames from the UCF50 dataset, which contains a total of 50 action classes across real-world scenarios. The images represent example frames from five selected classes, such as Basketball, Floor Fall, Push, Walk and Horse Race.

Fig. 4. Sample video frames from HMDB51 dataset, which contains 51 action classes from different action classes captured from different real-world and movie scenarios. The images represent example frames from five selected classes, including Catch, Pull Up, Laugh, Pour, and Turn.

variations, while ReLu non-linearities minimize sensitivity to low-light artifacts. Next, the MRSBU module suppresses noise, including inconsistencies in lighting and minor deviations in posture, using learnable soft-thresholding where channel-wise thresholds adaptively filter non-discriminative features during residual refinement. This process makes handcrafted pre-processing unnecessary, as the model inherently suppresses noise. Raw video frames are pre-processed and resized using the Inception v3 module to ensure uniform input dimensions before the feature extraction phase.

Characteristic	UCF101	UCF50	HMDB51
Clip Duration	1–10 s	1–10 s	1–10 s
Diversity	High (sports, acts)	Moderate (subset of UCF101)	High (natural, expressive actions)
Noise Level	Low to Moderate	Low to Mederate	High compresssion,blur,
Availability	Freely available	Freely available	Freely available
Scene Background	Moderate	Moderate	Highly variable
Pose Variiability	Mostly consistent	Controlled	Highly variable
Number of Clips per Clas	~130	~132	~100–150

Fig. 5. Comparison of characteristics across UCF101, UCF50 and HMDB51 action recognition video dataset.

4. Proposed method

In this study, an MRSBU-GRU approach is developed for efficient HAR in videos. First, videos are converted into frames during the pre-processing phase. Then, significant features are extracted using Inception v3, which extracts the deep features to recognize the human activities. Finally, activities are classified using the MRSBU-GRU method, which achieves high recognition accuracy. Fig. 6 illustrates the process of human activity recognition in videos.

4.1. Feature extraction

Inception v3 is a DL-based method that provides highly reliable performance in feature extraction and classification tasks. The method employs Directed Acylic Graph (DAG) with 94 convolutional layers, 316 layers in total, and 350 connections (UCF50 dataset, 2025, HMDB51 dataset, 2025, Ramaneswaran et al., 2021). Inception v3 is the backbone of spatial feature extraction because of its factorized convolutional architecture and parallel multi-branch design, which offer superior representation for modeling complex spatiotemporal patterns in action recognition. Inception v3 extracts deep spatial features from video frames, enabling improved feature extraction by capturing both fine-grained and high-level semantic features essential for distinguishing subtle features in human activity. Unlike ResNet, which majorly increases layer depth to improve performance, Inception v3 used asymmetric convolutions ($1 \times n$ *followed by* $n \times 1$) and parallel filter processing ($1 \times 1,\ 3 \times 3,\ 5 \times 5$) to capture spatial data at multiple scales without increasing computational cost. This allows the model to process fine-grained visual features (e.g., facial orientation, hand movement) as well as global contextual features (e.g., background interactions, motion). These abilities are significant for differentiating similar classes of action frames. EfficientNet focuses on compound scaling, but its uniform architecture limits the diversity of spatial feature maps when

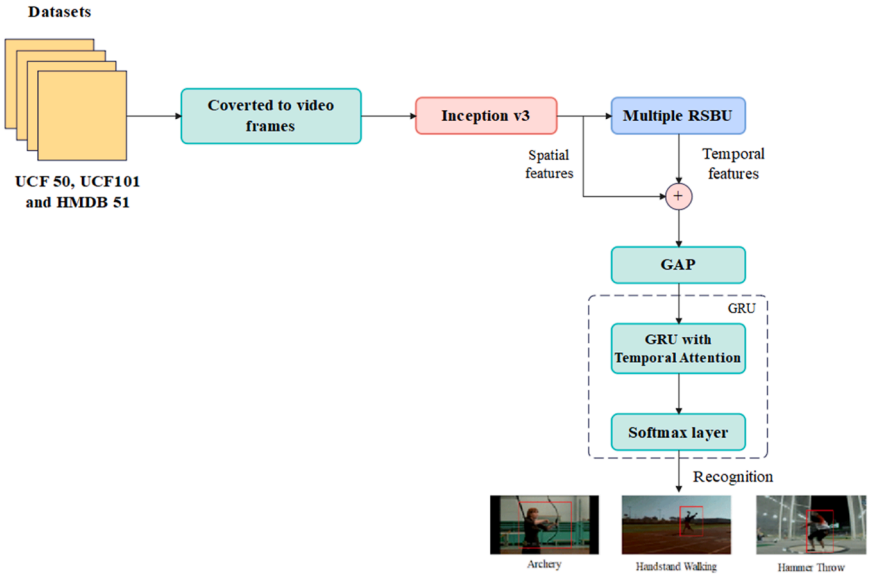

Fig. 6. Overall workflow of the proposed MRSBU-GRU model using the UCF and HMDB dataset.

applied to highly varied human activity frames. In contrast, Inception v3 uses 1×1 convolution for dimensionality reduction and auxiliary classifiers for regularization, thereby improving generalization in video-based tasks. Transfer learning is employed by initializing the network with ImageNet pre-trained weights. Early convolutional layers are frozen to retain low-level visual filters, while deep inception blocks— particularly those close to mixed and fully connected layers—are fine-tuned to adapt to the task-specific semantics of video dataset. This selective training strategy minimizes overfitting, accelerate convergence, and ensures that the model captures domain-relevant patterns.

The inception modules integrate multiple kernel sizes $(1 \times 1,\ 3 \times 3,\ 5 \times 5)$ in parallel, enabling simultaneous capture of intricate features. This minimizes computational cost while preserving representation capability. It also improves gradient flow, enhances feature learning in initial layers, and balances depth and multi-scale width, resulting in rich semantic feature representations. The DAG network architecture is challenging due to multiple inputs are being processed across various layers in a short span of time. To address this, numerous masks are introduced at different layers for capturing several features. Unlike traditional CNNs, which rely on predefined parameters, Inception v3 assigns masks and parameters across various layers. Inception v3 is trained on ImageNet, a challenging dataset containing millions of images across 1000 classes. The DL-based method learns from multiple objects and classes. ReLU activation and batch normalization are employed to introduce non-linearity and stabilize training. The equations for ReLU and pooling operations are provided in Appendix A.

Multiple layers are fused before applying an average pooling layer. Features are then captured from the average pooling layer. From the Inception v3 model, a total of 2048 features per image are extracted using Global Average Pooling (GAP) and fed into the classification phase. Table 2 presents the parameters of the Inception v3-based feature extraction model.

4.2. Human activity recognition

The Gated Attention Recurrent Network (GARN) is a hybrid mechanism that integrates a GRU and an attention mechanism. Several issues arise when using a Recurrent Neural Network (RNN), which assigns previous data instead of present data for classification. To mitigate this issue, GRU and the attention mechanism are integrated, forming the GARN. MRSBU redefines traditional residual learning by introducing a non-linear, adaptive soft-thresholding mechanism. Conventional residual blocks add learned feature transformations $F(x)$ to the input x, thereby allowing the network to preserve data and mitigate vanishing gradients. However, all features, including low-activation, noisy, and semantically inappropriate features, are transmitted forward without discrimination. In the proposed model, the residual learning is improved through the integration of MRSBU, which extends the traditional residual blocks by incorporating an adaptive soft-thresholding mechanism. Unlike standard residual connections that forward all feature activations, MRSBU selectively suppresses noisy and low-relevance features using attention-guided thresholds. This dynamic filtering improves feature sparsity and semantic relevance, allowing the network to focus on informative temporal patterns. By embedding this mechanism into the residual learning framework, MRSBU enhances gradient flow, improves robustness to noise, and strengthens generalization ability. MRSBU combines residual shrinkage with channel-wise adaptive and soft-thresholding, where τ_c is learned. This process ensures that features with sufficient activation magnitudes are retained, effectively acting as a sparsity filter embedded in the residual connection. The attention-guided thresholding in MRSBU improves adaptivity by making thresholds τ_c, a process of input activation statistics. The GAP is applied to every channel, generating a summary activation vector, which is fed into a lightweight fully connected attention subnetwork to produce scaling parameters. The thresholds are then calculated as $\tau_c = \alpha c$. This mechanism allows the network to dynamically contextualize threshold values based on the present input, resulting in selective feature suppression.

As a result, it mitigates noise and background artifacts while highlighting salient spatiotemporal features significant for human action recognition. To increase effectiveness, MRSBU is employed at multiple spatial levels during feature extraction. In early layers, it suppresses high-frequency noise and illumination artifacts. In intermediate layers, it filters out background variations, and in deep layers, it improves high-level semantic activations. This phase employed with information bottleneck principle, where early layers focus on entropy reduction and later layers improves task-specific relevance. This architecture provides better feature compression, enhanced gradient-flow and strong generalization, validating the necessity of adaptive thresholding across multiple processing phases. Fig. 7 represents the architecture of MRSBU-GRU model.

In Fig. 7, the detailed architecture of human activity recognition is presented by integrating spatial feature extraction, temporal modeling, and attention mechanisms. The proposed model considers a sequence of video frames, each of size 320 $\times$ 240 as input.

Table 2
Configuration parameters of the Inception v3 model used for spatial feature extraction in HAR.

Parameters	Value
Input Resolution	320×240
Normalization	Mean = [0.458, 0.456, 0.406] SD = [0.229, 0.224, 0.225]
Feature extraction layer	GAP (2045-D)
Weight Initialization	ImageNet pre-trained weights
Batch Normalization	Momentum = 0.99, δ=0.001
Optimizer	Adam $lr = 1e-4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$
Batch size	32
Activation function	ReLU

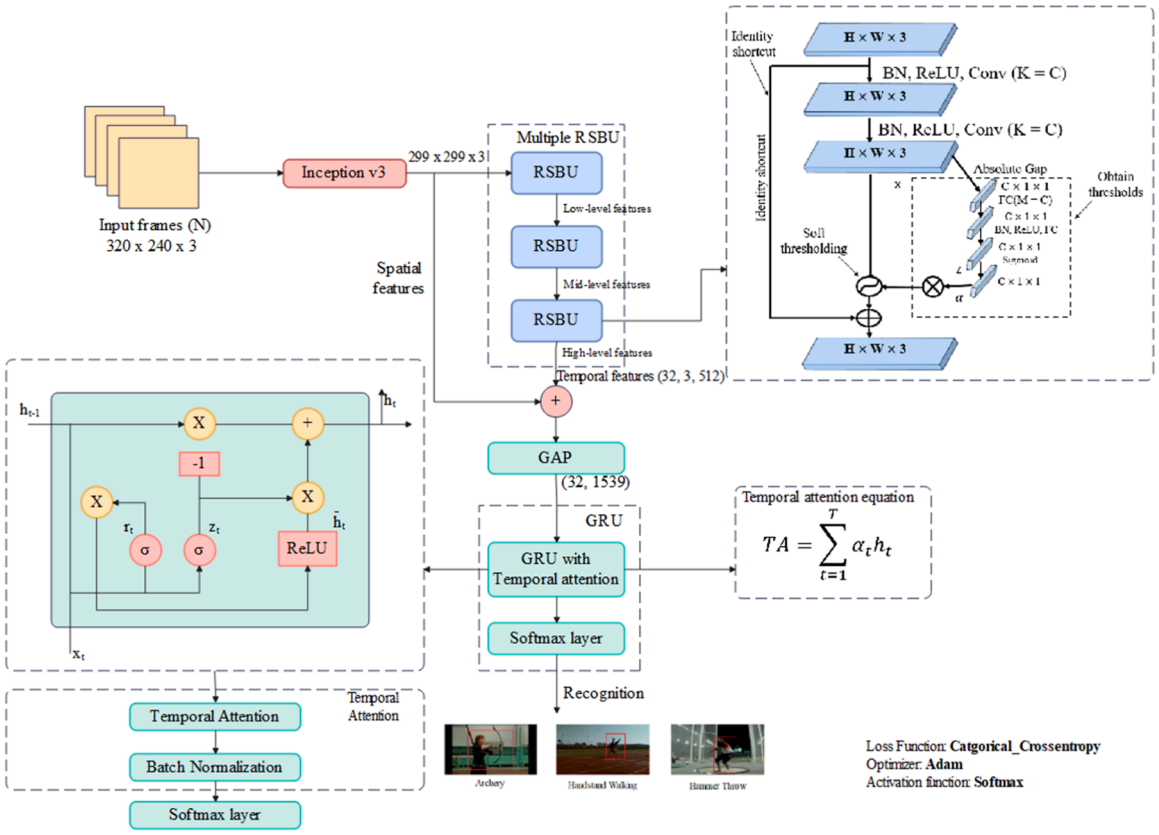

Fig. 7. Architecture of MRSBU-GRU model integrating Inception v3 for spatial feature extraction and RSBU for temporal feature learning.

These frames are initially processed by Inception v3, a CNN-based model utilized to capture rich spatial features, producing a feature map of size 299 $\times$ 299 $\times$ 3 for every frame. The spatial features, are then passed to multiple RSBUs that learn complex temporal relationships across the sequence while preserving spatial data. The RSBU blocks output a tensor of size (32,3, 512), where 32 is the batch size, 3 is the dimension, and 512 is the number of feature channels. A side module performs soft-thresholding using 1 $\times$1 convolution layers and ReLU activation, which helps to suppress noise and improve discriminative features by learning adaptive

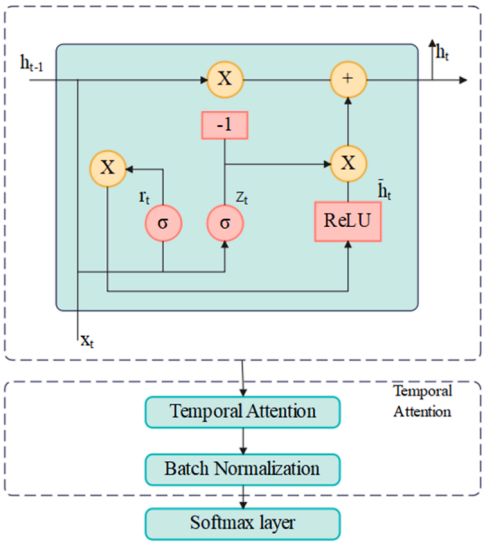

Fig. 8. Internal architecture of GRU cell with temporal attention for action recognition.

thresholds. The dimensionality of the refined features is reduced by GAP, resulting in a feature vector of size (32,1539) that summarizes temporal features across the frame sequence. This vector is then fed into the GRU with a temporal attention mechanism, which assigns dynamic significance to every frame by $c = \sum_{t=1}^{T} \alpha_t h_t$, where h_t is the hidden state and α_t is the learned attention weight for each time step t. Finally, the weighted context vector is fed into the softmax layer for recognition.

4.2.1. Gated recurrent unit (GRU)

To resolve the vanishing gradient issue, GRU is used in this research. GRU is considered an improved version of LSTM but has a null output gate. The GRU contains reset and update gates, which enable it to pass data forward across multiple time windows to produce accurate predictions or classifications. Specifically, weights and data are stored in memory are utilized with the given state to update values. The parameters of GRU considered in this manuscript are the categorical cross-entropy loss function, Adam optimizer, softmax activation function, batch size of 32, and 100 epochs. To mitigate overfitting, a dropout of 0.3 is applied after the GRU layer, and stratified sampling is utilized for dataset partitioning to preserve class distribution across training and validation sets. GRU incorporates update and reset gates to regulate data flow across time steps. The corresponding mathematical formulations are given in Appendix A. Fig. 8 presents the internal structure of a GRU cell with temporal attention. The update and reset gates regulate data flow, followed by attention-based weighting to highlight relevant time steps. The output is normalized by batch normalization and classified using the softmax layer.

4.2.2. Attention mechanism

The attention module is essential for determining the correlation between human activities. The attention method utilized here is a feed-forward attention method. In the proposed model, temporal attention is integrated into the GRU to improve its ability to model the complex temporal dynamics across video frames. While GRU efficiently captures sequential dependencies through its gated recurrent units, it traditionally compresses the entire input sequence into the final hidden state. This often leads to temporal data loss, particularly in long video sequences where action-related features occur. To mitigate this, a temporal attention module is introduced, which processes the sequence of hidden states produced by GRU. This module employs learned importance weights of every time step, enabling the model to dynamically prioritize frames containing semantically discriminative features (e.g., hand gestures, sudden motion) while down-weighting irrelevant features (e.g., background motion). The attention mechanism efficiently develops a temporal saliency map across the GRU outputs and guides the model to selectively aggregate meaningful data from the whole sequence. This improves the feature representation fed into the classification layer and improves the model's robustness to variable video lengths, motion sparsity, and intra-class variation. Moreover, it allows gradient features to propagate back across all time steps, enabling the attention-enhanced GRU to attain more stable training and better convergence. This process generates a single vector v from the entire sequence, as shown in Eqs. (10–12),

$$E_{t1} = b(H_{t1}) \tag{10}$$

$$\beta_{t1} = \frac{\exp(E_{t1})}{\sum_{s=1}^{R} \exp(E_s)} \tag{11}$$

$$v = \sum_{t_1=1}^{R} \beta_{t_1} H_{t_1} \tag{12}$$

$$TA = \sum_{t=1}^{T} \alpha_t h_t \tag{13}$$

In Eqs. (10–13), β represents the learning function, and is identified through H_{t1}, the α_t denotes the attention weight for time t. The attention mechanism generates a fixed-length vector for each sequence by calculating the weighted mean of the data sequence. For classification, the softmax layer is used to normalize outputs into class probabilities, and cross-entropy loss is employed for training. Detailed equations for softmax and cross-entropy are provided in Appendix A.

4.2.3. Shrinkage mechanism

When facing high levels of noise, Inception v3, which is used as a feature extractor, may fail to detect human activities. This results in reduced discrimination capacity of the high-level features learned through the output layer, which is insufficient for precise recognition of human activity. To address this, RSBU is developed. RSBU integrates ResNet's identical cross-layer connections, enhancing the model's stability and training effectiveness, while also employing soft-thresholding and attention mechanisms. RSBU is a module developed to capture temporal dependencies in a sequence of video frames. Each RSBU processes feature maps extracted from Inception v3 and integrates 3D convolutions, residual connections, and non-linear activations to capture motion patterns and frame-to-frame variations. The incorporation of residual connections ensures better gradient flow and stability during training, allowing the network to learn deeper representations. The RSBU consists of convolutional layers employed across spatial and temporal dimensions, enabling the extraction of fine-grained motion features and temporal structures. By stacking multiple RSBU blocks, the model builds a hierarchical understanding of spatiotemporal features across various abstraction levels. These blocks are essential for transforming initial frame-wise spatial features into rich temporal feature sequences, which can then be efficiently pooled and analyzed by subsequent GAP and GRU layers. Thus, RSBU plays an essential role in developing spatial CNN-based representations with

temporal sequence modeling, making the framework efficient for action recognition in videos. This approach also mitigates difficulties faced in feature extraction using convolutional kernels. Initially, the input data is fed into the RSBU module. The settings considered are $I(t) = x,\ Q(t) = y$ and $P(t) = z$. Within the RSBU, features are split into two branches, with one branch applying two convolutional layers for refinement. Each convolutional layer is followed by Batch Normalization (BN) and ReLU process. The mathematical formulation of this process is given in Eqs. (17) and (18):

$$x_j^k = f\left(\sum_{N}^{i=1} x_i^{k-1} \times w_{ij}^k + b_j^k\right) \tag{17}$$

$$f(x) = (0, x) \tag{18}$$

In Eqs. (17) and (18), x_i^{k-1} represents the i^{th} feature vector of the previous $(k-1)^{th}$ layer, x_j^k represents the j^{th} feature vector of the present k^{th} layer, and N denotes the number of input feature vectors. w_{ij}^k and b_j^k are the weights and bias of the neurons, respectively. The $\times$ represents the convolution process, and the $f(x)$ represents ReLU that maximizes non-linear expressiveness. After these two convolution layers, the acquired feature x is processed and reduced to a 1D vector using Global Average Pooling (GAP). Finally, feature A is acquired and the mathematical expression is given in Eq. (19):

$$A = average\left|x_{i,j,c}\right| \tag{19}$$

In Eq. (19), $i,\ j,\ and\ c$ represent the width, height, and channel matrices of feature x. This 1D vector is then separated into two paths. The inputs are passed through the attention mechanism, which includes two Fully Connected (FC) layers. The first FC layer includes BN and ReLU, while the second FC layer contains ReLU. The number of output neurons in this module matches the number of input feature channels. The mathematical formula for normalizing outputs between 0 and 1 is given in Eq. (20):

$$\alpha_c = \frac{1}{1 + \exp(-Z_c)} \tag{20}$$

In Eq. (20), Z_c represents a neuron, and α represents its respective scaling parameter. The mean value of different channels is multiplied by the result of the attention module for acquiring the threshold for each channel. Its mathematical formula is given in Eq. (21):

$$\tau_c = \alpha_c \times A \tag{21}$$

In Eq. (21), τ_c represents the threshold for channel c of feature x. Using the acquired outcomes, the soft- thresholding process is performed and then included to the actual features to obtain cross-layer identity connections. The mathematical formula for this process is given in Eq. (22):

$$M = m + \tau_c \times m,\ \ (M = X, Y, Z; m = x, y, z) \tag{22}$$

In Eq. (22), M represents the RSBU output as well as the input. $X,\ Y\ and\ Z$ are the results of $I-$ channel. After processing through multiple RSBU layers these features are fed into the GRU layer. At the output layer, the softmax activation function is used to convert raw output scores into probability values, enabling multi-class classification. This ensures that the sum of all output probabilities equals 1, and the class with the highest probability is chosen as the final prediction. This makes the model suitable to determine much likely class among multiple possible outcomes.

4.2.3.1. Pseudocode –MRSBU-GRU model for human activity recognition. Input –Dataset $V = \{v_1,\ v_2,\ \ldots, v_n\}$,
Output –Predicted labels Y_{pred}Frame processing and Feature extraction
For every video v_i in V:
$F_i \leftarrow extract\ features\ (v_i)$For every frame f_j in F_i:
$f_j \leftarrow resize\ and\ normalize\ f_j features\ [j] \leftarrow Inception\ v3\ \left(f_j\right)\ (2048 - dimension\ vector) Fv[i] \leftarrow features$Apply Multiple Residual Shrinkage Building Units (MRSBU)
For i from 1 *to* n:
sequence features $\leftarrow$[]For every frame feature f_j in $Fv[i]$:
$x \leftarrow Conv - BN - ReLU\ \left(f_j\right) x \leftarrow Conv - BN - ReLU\ (x)$Global Average Pooling
$gap \leftarrow GlobalAveragePooling\ (x)$Attention-based threshold (t) generation
$t \leftarrow FullyConnected(gap) t \leftarrow BatchNorm(t) t \leftarrow ReLU(t) t \leftarrow FullyConnected(t) t \leftarrow Sigmoid(t)$Calculated threshold per channel
$threshold \leftarrow t \times Mean(|x| across\ spatial\ dimensions)$Soft-thresholding (shrinkage)
For every element x_c in x:
If $abs(x_c) > threshold_c$:
$x_c \leftarrow sign(x_c) \times (abs(x_c) - threshold_c)$Else:
$x_c \leftarrow 0$Residual connection
$out \leftarrow x + f_j sequence\ features\ (out) s[i] \leftarrow sequence features$Temporal Modeling with GRU and Attention
For i from 1 to n:

$H \leftarrow GRU(s[i])$Attention weights
For every h_j in H:
$\alpha_j \leftarrow softmax\ (score(h_j))$Calculate context vector
$C \leftarrow 0$For j from 1 *to* k:
$C \leftarrow C + \alpha j \times h_j y_{hat} \leftarrow Softmax(FC(C))$If $\max(y_{hat}) > threshold$:
$y_i \leftarrow argmax(y_{hat})$Else:
$y_i \leftarrow Uncertain$Return Y_{pred}

4.2.3.2. Research implications. The findings of this research have several essential implications for HAR and computer vision. By combining MRSBU with GRU, the developed method demonstrates a significant advancement in effectively capturing both spatial and temporal dynamics in video data. This method improves recognition accuracy on standard dataset and enhances robustness in real-world, unconstrained environments. The strong generalization ability of the model suggests practical applicability in critical domains such as surveillance, healthcare monitoring, and human-computer interaction.

5. Experimental analysis

The developed algorithm is evaluated using the performance measures of accuracy (overall performance), precision (positive prediction accuracy), recall (sensitivity or coverage of original positive samples), and F1-score (mean average of precision and recall). The mathematical formulae for these performance measures are given in Eqs. (23,24,25,26),

$$Accurcay = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

$$Precision = \frac{TP}{TP + FP} \tag{24}$$

$$Recall = \frac{TP}{TP + FN} \tag{25}$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \tag{26}$$

5.1. Implementation details

The dataset is partitioned into 80 % for training, 10 % for testing, and 10 % for validation to ensure robust model evaluation. Table 3 presents the system configurations used for model training and evaluation, while Table 4 presents the training hyper-parameters and configuration settings utilized in the proposed MRSBU-GRU model. The proposed model is trained using the Adam optimizer, because of its adaptive learning rate and fast convergence on sparse gradients present in temporal and spatial feature learning. The initial learning rate of 0.001 is chosen to provide balance trade-off between convergence speed and stability. To prevent overfitting and ensure efficient convergence, the ReduceLROnPlateau learning rate scheduler is used. This minimized the learning rate by a factor of 0.1 when validation loss does not improve for 5 consecutive epochs, allowing the model to fine-tune its learning in later phases. The categorical cross-entropy loss is used, as it is standard for multi-class classification. The ReLU activation function is applied to hidden layers to introduce non-linearity and avoid vanishing gradients, while the Softmax is used at the output layer to convert outputs into class probabilities for multi-class classification. The batch size of 32 is chosen to ensure sufficient gradient estimation without draining GRU memory. The model is trained for 100 epochs, which provides enough time to capture spatial and temporal dynamics. Each epoch requires approximately 50 s, reflecting the model complexity and GPU acceleration utilized, that aligns with the system configuration and training strategy. A dropout rate of 0.3 is set, introducing moderate regularization to reduce overfitting without hindering learning. A high dropout rate would cause underfitting and slow convergence, whereas a low dropout rate would fail to sufficiently prevent co-adaption of neurons.

When accuracy is a widely used metric in recognition tasks, it does not completely capture model performance, particularly in real-world scenarios such as anomaly detection. Hence, in addition to accuracy, the proposed model is evaluated using precision, recall, F1-

Table 3
System configuration used for model training and evaluation.

Components	Value
RAM	16 GB
Processor	Intel i5
Operation-OS	Windows 10
GPU	6 GB
SSD	1 TB

Table 4
Training hyperparameters and configuration settings used on proposed MRSBU-GRU model.

Parameter	Value
Optimizer	Adaptive Moment Estimation (Adam)
Initial learning rate	0.001
Learning rate scheduler	ReduceLROnPlateau (factor = 0.1, patience = 5)
Loss Function	Categorical Crossentropy
Activation functions	ReLU (hidden layers), Softmax (output layer)
Batch size	32
Number of epochs	100
Training time per epoch	50 s
Dropout rate	0.3

score, and AUC to provide a more understanding of its effectiveness. In Table 5, the performance of feature extraction using Inception v3 is depicted using UCF50, UCF101, and HMDB51 dataset. EfficientNet, MobileNet, DenseNet, VGG16, and ResNet are considered as existing algorithms to validate the performance of Inception v3. The inception modules integrate various kernel sizes $(1\times1,\ 3\times3,\ 5\times5)$ in parallel, enabling the simultaneous capture of intricate features. This minimizes computational cost while preserving representational ability. It also improves gradient flow, enhances feature learning in the initial layers, and balances depth with multi-scale width, finally delivering rich semantic feature representations. Inception v3 is chosen as the backbone for spatial feature extraction because of its ability to capture multi-scale patterns using factorized convolutions while maintaining reasonable computational efficiency. In comparison, EfficientNet or ResNet, which are optimized for image classification, but are computationally demanding for spatiotemporal contexts. Inception v3 provides a balanced trade-off between accuracy and efficiency. Lightweight models like MobileNet offer minimized complexity but are less efficient in preserving rich spatial data required for activity recognition. The Inception v3-based feature extraction algorithm achieves an accuracy of 99.75 % on the UCF50 dataset, 99.55 % on the UCF101 dataset, and 98.95 % on the HMDB51 dataset.

In Table 6, performance of proposed algorithm without shrinkage is evaluated using UCF50, UCF101 and HMDB51 dataset. In the below Table 4, Gated Recurrent Unit (GRU), Recurrent Neural Network (RNN), CNN-LSTM, LSTM, and CNN are evaluated without shrinkage mechanism.

In Table 7, the performance of the proposed algorithm with the shrinkage mechanism is evaluated using UCF50, UCF101, and HMDB51 dataset. Here, different classifiers such as GRU, RNN, CNN-LSTM, LSTM, CNN with shrinkage mechanism, VideoMAE, and TimeSformer are evaluated. While CNN, RNN, and LSTM variants capture local features, they struggle with long-range temporal dependencies. Transformer-based models like VideoMAE and TimeSformer require large amounts of data and lead to overfit on small dataset. The proposed model uses MRSBU for adaptive noise suppression and Gated GRU for effective temporal modeling, resulting in robust and discriminative representations. The proposed algorithm achieves an accuracy of 99.75 % on the UCF50 dataset, 99.55 % on the UCF101 dataset, and 98.95 % on the HMDB51 dataset.

To analyze the effectiveness of the attention mechanism within the proposed MRSBU-GRU model, evaluations are conducted across five benchmark dataset, as presented in Table 8. The results show consistent performance when the attention model is integrated. The proposed model achieves accuracy improvements of 1.40 %, 1.95 %, and 2.15 % for UCF50, UCF101, and HMDB dataset, respectively. More significant improvements are obtained on real-world dataset, with 2.15 % on UCF-Crime and 3.98 % on UCSD dataset, where noise and background variability are more prominent. These improvements highlight the role of the attention module in improving spatial feature relevance by highlighting informative channels and suppressing redundant activations. This, in turn, provides more

Table 5
Performance of Inception v3 based feature extraction algorithm using UCF50, UCF101 and HMDB51 dataset.

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EfficientNet	UCF 50	91.45	90.80	91.10	90.95
	UCF 101	89.75	88.90	88.40	88.65
	HMDB 51	87.55	86.70	87.10	86.90
MobileNet	UCF 50	93.85	92.40	93.20	92.80
	UCF 101	91.25	90.10	90.75	90.42
	HMDB 51	89.15	88.00	88.45	88.22
DenseNet	UCF 50	94.30	93.55	94.10	93.82
	UCF 101	92.95	91.70	92.30	92.00
	HMDB 51	90.85	89.60	89.95	89.77
VGG19	UCF 50	95.10	94.40	94.70	94.55
	UCF 101	93.65	92.85	92.50	92.67
	HMDB 51	91.35	90.25	90.60	90.42
ResNet	UCF 50	97.65	97.10	97.30	97.20
	UCF 101	96.90	96.35	96.50	96.42
	HMDB 51	95.75	95.20	95.40	95.30
Inception v3	**UCF 50**	**99.75**	**99.85**	**99.75**	**99.80**
	UCF 101	**99.55**	**99.40**	**99.40**	**99.40**
	HMDB 51	**98.95**	**99.30**	**98.20**	**98.75**

Table 6
Performance of the proposed algorithm without a shrinkage mechanism using UCF50, UCF101, and HMDB51 dataset.

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
GRU	UCF50	93.55	92.80	92.90	92.85
	UCF101	92.30	91.50	91.70	91.60
	HMDB51	91.10	90.70	90.50	90.60
RNN	UCF50	94.10	93.50	93.60	93.55
	UCF101	93.80	93.20	93.30	93.25
	HMDB51	92.70	92.10	92.00	92.05
CNN-LSTM	UCF50	96.25	95.90	96.10	96.00
	UCF101	95.60	95.40	95.50	95.45
	HMDB51	94.80	94.20	94.30	94.25
ViT	UCF50	97.00	96.80	96.90	96.85
	UCF101	96.40	96.10	96.20	96.15
	HMDB51	95.20	94.80	94.90	94.85
Gated GRU	UCF50	98.55	98.60	98.20	98.40
	UCF101	97.75	97.65	97.25	97.45
	HMDB51	96.90	96.75	96.30	96.52

Table 7
Performance of the proposed algorithm with a shrinkage mechanism using UCF50, UCF101, and HMDB51 dataset.

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Multiple Residual Shrinkage Building Units CNN	UCF 50	98.10	98.30	97.90	98.10
	UCF 101	97.45	97.25	96.85	97.05
	HMDB 51	96.50	96.30	95.85	96.07
Multiple Residual Shrinkage Building Units RNN	UCF 50	97.95	98.10	97.70	97.90
	UCF 101	97.10	97.00	96.45	96.72
	HMDB 51	96.25	96.00	95.70	95.85
Multiple Residual Shrinkage Building Units CNN-LSTM	UCF 50	98.35	98.40	98.10	98.25
	UCF 101	97.60	97.50	97.00	97.25
	HMDB 51	96.80	96.65	96.10	96.37
Multiple Residual Shrinkage Building Units ViT	UCF 50	98.00	98.05	97.80	97.92
	UCF 101	97.25	97.20	96.60	96.90
	HMDB 51	96.00	95.80	95.50	95.65
Multiple Residual Shrinkage Building Units GRU	UCF 50	98.55	98.60	98.20	98.40
	UCF 101	97.75	97.65	97.25	97.45
	HMDB 51	96.90	96.75	96.30	96.52
VideoMAE	UCF 50	96.82	96.55	96.68	96.61
	UCF 101	95.28	94.88	95.95	95.41
	HMDB 51	96.50	96.00	96.00	96.00
TimeSformer	UCF 50	94.00	94.21	94.47	94.34
	UCF 101	97.75	97.45	97.65	97.55
	HMDB 51	97.00	96.88	96.73	96.80
Proposed Multiple Residual Shrinkage Building Units Gated GRU	UCF 50	**99.75**	**98.85**	**99.75**	**99.80**
	UCF 101	**99.55**	**99.40**	**99.40**	**99.40**
	HMDB 51	**98.95**	**99.30**	**98.20**	**98.75**

Table 8
Proposed model performance with and without attention mechanism across multiple HAR dataset.

Dataset	Accuracy without attention (%)	Accuracy with Attention (%)	Improvement (%)
UCF50	98.35	99.75	1.40
UCF101	97.60	99.55	1.95
HMDB51	96.80	98.95	2.15
UCF Crime	97.80	99.95	2.15
UCSD	96.00	99.98	3.98

robust temporal modeling in complex activity scenarios, resulting in better generalization and recognition accuracy.

To evaluate the stability and generalization ability of the proposed MRSBU-GRU model, k-fold cross-validation is conducted with $k = 2$, 3, and 5 across benchmark dataset, as presented in Table 9. The proposed model achieves the highest performance with 5-fold validation, highlighting its strong generalization across different data splits. This improvement with higher fold values is contributed to training and testing the model on more varied subsets of the dataset, which minimizes the risk of overfitting and ensures a robust performance evaluation. These results confirm the proposed model's reliability and its ability to maintain high performance under various validation strategies.

To analyze the efficiency and statistical significance of the proposed MRSBU-GRU model, comparative analysis is performed against

Table 9
Performance evaluation of proposed model using different k-fold cross-validation using UCF 50, UCF 101 and HMDB 51 dataset.

K-folds	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
2	UCF 50	99.40	99.55	99.35	99.45
	UCF 101	99.10	98.95	98.90	98.90
	HMDB 51	98.10	98.50	97.80	98.10
3	UCF 50	99.60	99.70	99.55	99.60
	UCF 101	99.30	99.20	99.10	99.15
	HMDB 51	98.60	99.00	98.00	98.50
5	**UCF 50**	**99.75**	**99.85**	**99.75**	**99.80**
	UCF 101	**99.55**	**99.40**	**99.40**	**99.40**
	HMDB 51	**98.95**	**99.30**	**98.20**	**98.75**

other architectural variants, including CNN, RNN, CNN-LSTM, ViT, and standard GRU configurations, as presented in Table 10. Evaluations are performed across different benchmark dataset and consider performance measures such as p-value from *t*-tests, execution time, memory consumption, inference time, and training time. The proposed model consistently demonstrates superior performance in terms of computational efficiency and statistical robustness. These enhancements are obtained to the soft-thresholding mechanism in residual shrinkage blocks, which minimizes redundant spatial features, while GRU effectively captures temporal dependencies with fewer parameters and faster convergence compared to LSTM models.

To evaluate the computational complexity and scalability of different model configurations, evaluations are performed in terms of trainable parameters and Floating-Point Operations per Video (FLOPs). Table 11 presents the performance of MRSBU combined with six different integrated models using various temporal backbones. Among these architectures, the MRSBU + GRU provides a moderate number of parameters and balanced FLOPs, achieving an optimal trade-off between representation ability and computational cost. The proposed model configuration is chosen for its balance between accuracy, efficiency, and scalability, making it suitable for real-time applications where bot performance and resource constraints are essential.

5.2. Cross-dataset validation

To evaluate the generalization ability and cross-domain robustness of the proposed MRSBU-GRU model, two cross- dataset

Table 10
Performance of proposed model against baseline models across multiple metrics including statistical significance, computational efficiency and memory utilization.

Methods	Dataset	P value from *t*-test	Execution Time (s)	Memory Consumption (MB)	Inference Time (ms)	Traning Time (s)
Multiple Residual Shrinkage Building Units - CNN	UCF 50	0.055	2025	610	1340	1880
	UCF101	0.051	1910	623	1395	1725
	HMDB 51	0.083	2095	648	1280	1698
Multiple Residual Shrinkage Building Units - RNN	UCF 50	0.043	1895	595	1290	1620
	UCF101	0.047	1762	607	1322	1495
	HMDB 51	0.051	1950	635	1215	1457
Multiple Residual Shrinkage Building Units CNN-LSTM	UCF 50	0.036	1810	580	1260	1584
	UCF101	0.039	1705	592	1305	1436
	HMDB 51	0.042	1882	615	1187	1382
Multiple Residual Shrinkage Building Units ViT	UCF 50	0.031	1780	570	1235	1550
	UCF101	0.034	1650	585	1270	1392
	HMDB 51	0.037	1835	608	1152	1340
		0.055	2025	610	1340	1880
Multiple Residual Shrinkage Building Units GRU	UCF 50	0.041	1910	623	1395	1725
	UCF101	0.033	2095	648	1280	1698
	HMDB 51	0.021	1552	565	1205	1520
Proposed Multiple Residual Shrinkage Building Units Gated GRU	UCF 50	0.022	1622	578	1230	1140
	UCF101	0.028	1829	605	1120	1125
	HMDB 51	0.055	2025	610	1340	1880

Table 11
Comparison of trainable parameters and computational cost (FLOPs per video) for various MRSBU-integrated model architectures.

Models	Trainable Parameters (M)	FLOPs per Video (G)
MRSBU + CNN	25.3	8.1
MRSBU + RNN	30.1	7.25
MRSBU + LSTM	29.4	6.5
MRSBU + GRU	31.2	7.1
MRSBU + GRU+ViT	33.6	6.8
MRSBU + CNN-LSTM	28.9	6.15

experiments are conducted: training on UCF 101 and testing on UCF 50, and training on UCF 50 and testing on UCF 101. These results, presented in Tables 12 and 13, consistently show that the proposed model outperforms all baseline architectures. This demonstrates the effectiveness of the gating mechanism in capturing temporal dynamics and eliminating irrelevant features. The consistent performance across dataset shows that the proposed model learns spatial-temporal features and maintains high discriminative power when employed to various dataset.

To validate consistency, robustness, and performance stability, multiple experimental runs are performed, and key performance metrics are evaluated across different benchmark dataset. As shown in Table 14, the proposed model effectively outperforms all baseline architectures in terms of lower variance across multiple runs, representing high reliability. Baseline models demonstrated lower average metrics and higher variability, which leads to noise. The superior performance is obtained to the attention-guided soft-thresholding mechanism in residual shrinkage units, which suppresses inappropriate spatial activations, and the GRU, which effectively captures long-range temporal dependencies with lower computational overhead compared to baseline models. The lower standard deviations across all dataset confirm that the proposed model maintains generalization and consistency, making it suitable for real-world HAR that demand accuracy and stability.

Figs. 9, 15, and 21 present the class-wise accuracy of the UCF50, UCF101, and HMDB dataset, respectively. Figs. 10, 16, and 22 present the class-wise precision of the UCF50, UCF101, and HMDB dataset, respectively. Figs. 11, 17, and 23 present the accuracy graphs of the UCF50, UCF101, and HMDB dataset, respectively. Figs. 12, 18, and 24 present the loss graph of UCF50, UCF101, and HMDB dataset, respectively. Figs. 13, 18, and 24 present the confusion matrix of UCF50, UCF101, and HMDB dataset, respectively. Figs. 14, 20, and 26 present the ROC Curves of UCF50, UCF101, and HMDB dataset, respectively. To address overfitting observed in training and validation accuracy and loss curves, several strategies are adopted to ensure that the model is trained and validated across different subsets of the dataset. This prevents the model from overfitting to specific training partitions and provides a more reliable evaluation of generalization performance. To overcome imbalance, in class distributions especially in dataset like UCF50 and UCF101, stratified sampling is employed during training to ensure that each mini-batch and fold maintains a representative ratio of all classes. At the model level, a dropout rate of 0.3 is applied after the GRU layer, which randomly disables a faction of neurons during training to prevent the model from relying excessively on specific paths through network. The use of dropout minimizes fluctuations in validation loss, while stratified sampling mitigates class imbalance effects, leading to stable training curves. Moreover, the combination of

Table 12
Cross dataset validation using training on UCF101 and testing on UCF 50 dataset.

Models	Metrics	UCF 50	UCF 101
Multiple Residual Shrinkage Building Units - CNN	Accuracy	95.12	94.86
	Precision	94.93	94.62
	Recall	94.75	94.35
	F1-Score	94.83	94.47
Multiple Residual Shrinkage Building Units - RNN	Accuracy	95.90	95.67
	Precision	95.62	95.30
	Recall	95.34	95.05
	F1-Score	95.48	95.17
Multiple Residual Shrinkage Building Units CNN-LSTM	Accuracy	96.45	96.24
	Precision	96.21	95.94
	Recall	95.98	95.70
	F1-Score	96.09	95.82
Multiple Residual Shrinkage Building Units ViT	Accuracy	97.10	96.88
	Precision	96.87	96.61
	Recall	96.59	96.37
	F1-Score	96.73	96.49
Multiple Residual Shrinkage Building Units GRU	Accuracy	95.12	94.86
	Precision	94.93	94.62
	Recall	94.75	94.35
	F1-Score	94.83	94.47
Proposed Multiple Residual Shrinkage Building Units Gated GRU	Accuracy	97.98	98.65
	Precision	97.97	97.85
	Recall	97.98	97.76
	F1-Score	97.97	97.80

Table 13
Cross dataset validation using training on UCF50 and testing on UCF 101 dataset.

Models	Metrics	UCF 50	UCF 101
Multiple Residual Shrinkage Building Units - CNN	Accuracy	95.41	94.58
	Precision	94.87	94.12
	Recall	95.14	94.34
	F1-Score	96.24	95.35
Multiple Residual Shrinkage Building Units - RNN	Accuracy	95.88	95.07
	Precision	95.36	94.79
	Recall	95.61	94.93
	F1-Score	96.78	95.92
Multiple Residual Shrinkage Building Units CNN-LSTM	Accuracy	96.39	95.66
	Precision	96.01	95.37
	Recall	96.20	95.51
	F1-Score	97.36	96.42
Multiple Residual Shrinkage Building Units ViT	Accuracy	96.93	96.21
	Precision	96.47	95.94
	Recall	96.69	96.07
	F1-Score	95.41	94.58
Multiple Residual Shrinkage Building Units GRU	Accuracy	97.10	96.60
	Precision	96.80	96.35
	Recall	96.95	96.45
	F1-Score	97.00	96.40
Proposed Multiple Residual Shrinkage Building Units Gated GRU	Accuracy	98.00	97.00
	Precision	97.58	97.25
	Recall	97.00	97.15
	F1-Score	**97.29**	**97.20**

Table 14
Performance of proposed model with standard deviation across benchmark dataset for various MRSBU-integrated models.

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Multiple Residual Shrinkage Building Units - CNN	UCF 50	98.10 ± 0.15	98.30 ± 0.12	97.90 ± 0.18	98.10 ± 0.13
	UCF 101	97.45 ± 0.17	97.25 ± 0.14	96.85 ± 0.16	97.05 ± 0.15
	HMDB 51	96.50 ± 0.21	96.30 ± 0.19	95.85 ± 0.22	96.07 ± 0.20
Multiple Residual Shrinkage Building Units - RNN	UCF 50	97.95 ± 0.16	98.10 ± 0.14	97.70 ± 0.17	97.90 ± 0.13
	UCF 101	97.10 ± 0.19	97.00 ± 0.15	96.45 ± 0.18	96.72 ± 0.14
	HMDB 51	96.25 ± 0.22	96.00 ± 0.20	95.70 ± 0.24	95.85 ± 0.19
Multiple Residual Shrinkage Building Units CNN-LSTM	UCF 50	98.35 ± 0.13	98.40 ± 0.12	98.10 ± 0.15	98.25 ± 0.11
	UCF 101	97.60 ± 0.15	97.50 ± 0.14	97.00 ± 0.16	97.25 ± 0.13
	HMDB 51	96.80 ± 0.19	96.65 ± 0.18	96.10 ± 0.21	96.37 ± 0.17
Multiple Residual Shrinkage Building Units ViT	UCF 50	98.00 ± 0.14	98.05 ± 0.13	97.80 ± 0.16	97.92 ± 0.12
	UCF 101	97.25 ± 0.16	97.20 ± 0.15	96.60 ± 0.17	96.90 ± 0.14
	HMDB 51	96.00 ± 0.20	95.80 ± 0.19	95.50 ± 0.22	95.65 ± 0.18
Multiple Residual Shrinkage Building Units GRU	UCF 50	98.55 ± 0.12	98.60 ± 0.11	98.20 ± 0.14	98.40 ± 0.10
	UCF 101	97.75 ± 0.14	97.65 ± 0.13	97.25 ± 0.15	97.45 ± 0.12
	HMDB 51	96.90 ± 0.18	96.75 ± 0.17	96.30 ± 0.20	96.52 ± 0.16
Proposed Multiple Residual Shrinkage Building Units Gated GRU	UCF 50	99.75 ± 0.08	99.85 ± 0.07	99.75 ± 0.09	99.80 ± 0.06
	UCF 101	99.55 ± 0.09	99.40 ± 0.08	99.40 ± 0.10	99.40 ± 0.07
	HMDB 51	98.95 ± 0.11	99.30 ± 0.10	98.20 ± 0.13	98.75 ± 0.09

attention mechanism allows the model to selectively focus on the most relevant temporal frames, minimizing the impact of noisy and redundant data. This improves feature discrimination and acts as a form of regularization by dynamically adjusting the model's sensitivity to input sequences. The integrated model effectively minimizes the gap between training and validation performance, leading to more stable and generalizable outcomes across all dataset.

Figs. 19, 25

5.3. Analysis of generalization

For evaluating generalization, the performance of the proposed algorithm is tested on two additional dataset, namely UCF-Crime and UCSD, which are video dataset containing crime-related content. Compared to other dataset, the proposed algorithm delivers greater performance compared to existing algorithms. The generalization ability of the MRSBU-GRU method is assessed to ensure robustness across varied and real-world scenarios beyond benchmark dataset. While the method demonstrated performance on dataset such as UCF50, UCF101, and HMDB51, further validation is conducted using challenging dataset like UCF-Crime and UCSD, which include complex, unconstrained environments and abnormal human behaviors. These dataset are characterized by essential variability in video quality, camera motion, lighting conditions, background clutter, and presence of rare or anomalous activities, testing the

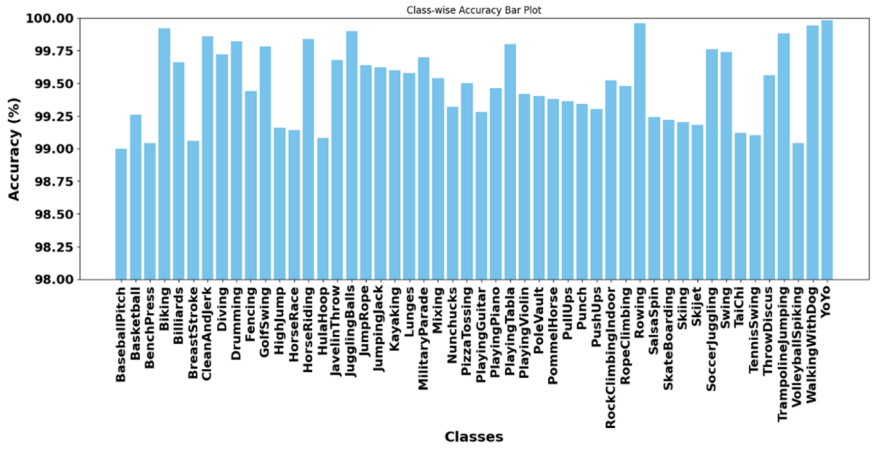

Fig. 9. Class wise accuracy of proposed model across UCF50 dataset action classes.

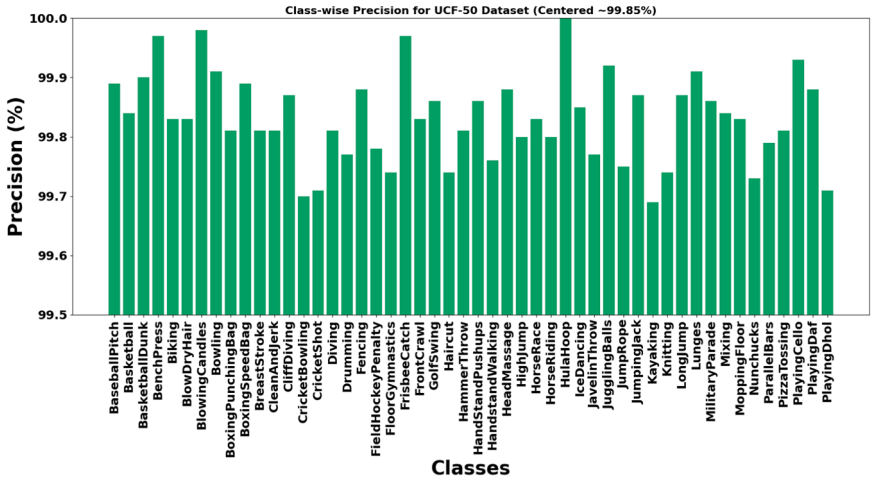

Fig. 10. Class-wise precision of proposed model across UCF50 dataset action classes.

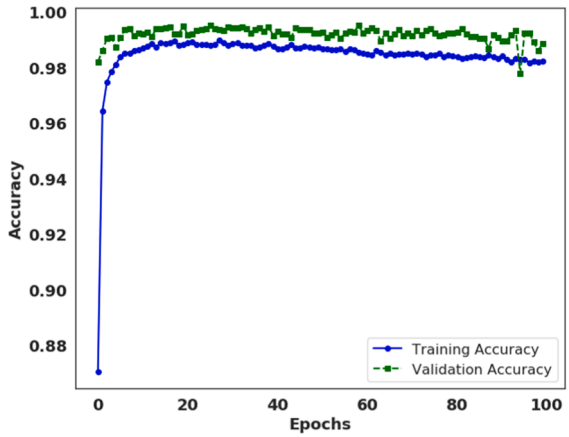

Fig. 11. Accuracy vs Epochs for proposed model across UCF50 dataset.

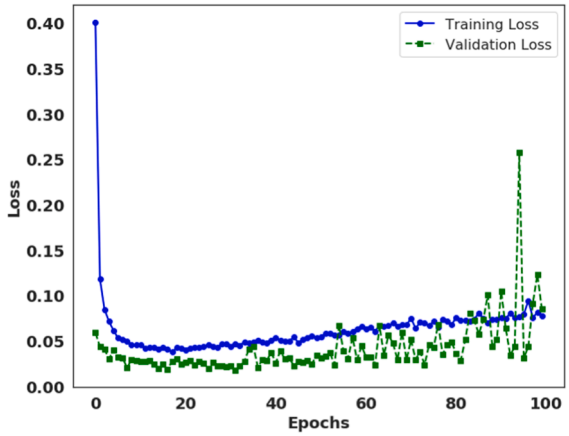

Fig. 12. Loss vs Epochs for proposed model across UCF50 dataset.

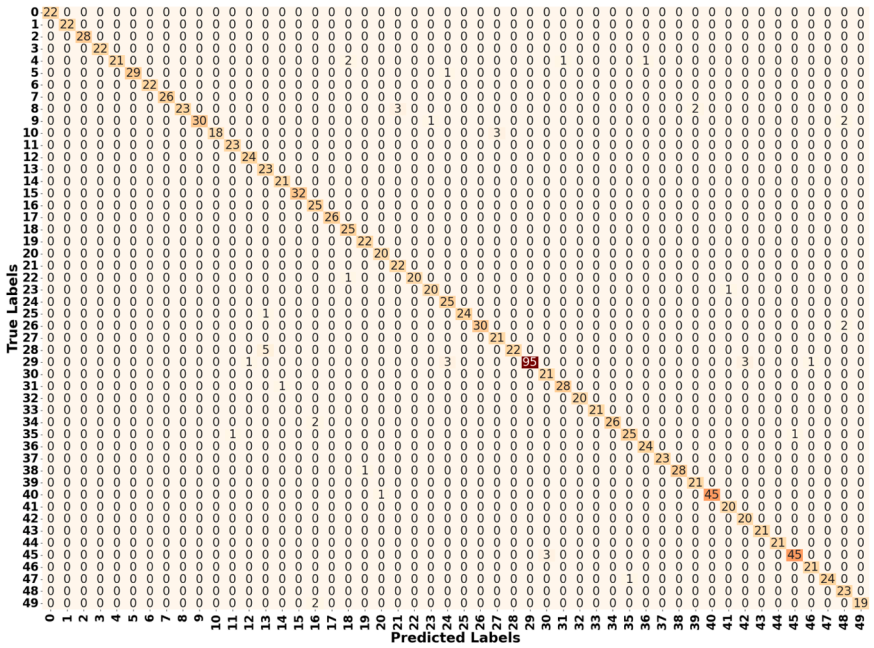

Fig. 13. Confusion matrix for proposed model across UCF 50 dataset.

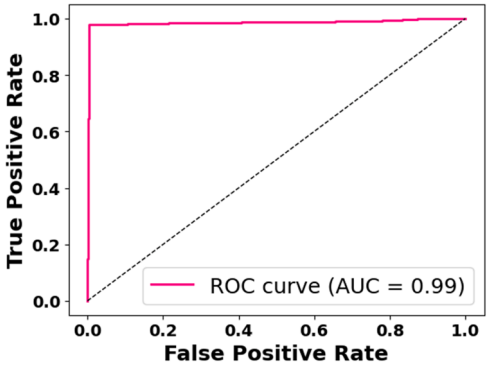

Fig. 14. ROC Curve of proposed model across UCF50 dataset.

model's resilience to distribution shifts. Despite these challenges, the MRSBU-GRU method achieves 99.95 % accuracy on UCF-Crime and 99.98 % accuracy on UCSD, effectively outperforming other variants and conventional methods. This strong generalization performance is attributed to the method's ability to suppress inappropriate or noisy features through dynamic soft-thresholding in MRSBU, along with efficient temporal modeling through GRU and attention mechanisms. These results demonstrate that the developed method performes efficiently in both controlled experimental setups and real-world video surveillance and behavioral analysis contexts, showing its practical applicability and robustness. Table 15 presents the generalization ability of the proposed algorithm.

5.4. Ablation study

To validate the individual and combined contributions of key components in the proposed model, an ablation study is performed using different configurations across benchmark dataset, as presented in Table 16. Baseline models such as Inception v3 + GRU lack adaptive feature suppression, which limits their accuracy. Incorporating attention into GRU + Attention enhances the model's ability to focus on salient temporal features, resulting in improved precision and F1-score across all dataset. The incorporation of MRSBU + GRU further increases performance by improving spatial feature robustness through soft-thresholding and residual connections. The better performance is obtained when attention is incorporated with MRSBU and GRU, resulting in high accuracy across all dataset. These results show that the effects of spatial attention, residual shrinkage, and gated temporal modeling effectively improve the model's ability to generalize across all video dataset. The improvements in F1-score, especially for HMDB51, determine the model's effectiveness in handling noisy and unstructured video data.

5.5. Comparative analysis

In this section, the performance of Multiple Residual Shrinkage Building Units integrated with GRU is compared to existing algorithms using UCF50, UCF101, and HMDB51 dataset. Table 17 represents the comparative analysis using the UCF50 dataset. Table 18 represents the comparative analysis using the UCF101 dataset. Table 19 represents the comparative analysis using the HMDB51 dataset. The developed algorithm achieves an accuracy of 99.75 % on the UCF50 dataset, 99.55 % on the UCF101 dataset, and 98.95 % on the HMDB51 dataset, which are higher compared to existing algorithms. To validate the efficacy of the developed MRSBU-GRU method, comparisons are made with conventional DL algorithms and other existing algorithms. Baseline results from methods such as ViT-ReT (Wensel et al., 2023), HyEx-Net (Khan et al., 2024), Fine-tuned VideoMAE (Gupta & Mahapatra, 2025), and ConvST-LSTM-Net (Jayamohan & Yuvaraj, 2025) could not be directly replicated under identical pre-processing and evaluation criteria due to computational overhead and unavailable pre-trained weights. Additionally, recent state-of-the-art relying on large-scale transformers could not be included because of resource constraints. Nevertheless, comparisons are performed on the same dataset under similar conditions to ensure fairness. State-of-the-art models dependent on large language or vision-language methods are excluded, as the primary focus of this study is on lightweight architectures with high generalization performance.

6. Discussion

This research differs from existing HAR algorithms by integrating an adaptive residual shrinkage mechanism with a temporal attention-guided GRU, addressing the limitations of previous models. The integration of adaptive residual shrinkage ensures that noisy or redundant spatial features are dynamically suppressed before entering the temporal modeling phase. This spatial-level feature refinement enhances the discriminative power of temporal representations.The temporal attention-guided GRU assigns varying importance weights to different frames in a sequence, enabling the model to focus on semantically essential moments while down-weighting irrelevant features. The proposed MRSBU-GRU model demonstrates strong real-world applicability by combining high accuracy with computational efficiency. In video surveillance, the model reliably detects suspicious and abnormal human behaviors even in noisy, cluttered, or low-light environments, making it suitable for crime prevention. In healthcare monitoring, its ability to capture subtle motion transitions allows accurate detection of falls, unusual movements, or patient activity patterns. Its design also

Table 15
Generalization ability of proposed algorithm using UCF Crime and UCSD dataset.

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Multiple Residual Shrinkage Building Units with CNN	UCF Crime	96.10	95.75	96.00	95.87
	UCSD	94.30	94.10	94.20	94.15
Multiple Residual Shrinkage Building Units with RNN	UCF Crime	96.80	96.40	96.60	96.50
	UCSD	94.90	94.70	94.80	94.75
Multiple Residual Shrinkage Building Units with CNN-LSTM	UCF Crime	97.20	96.90	97.10	97.00
	UCSD	95.30	95.10	95.20	95.15
Multiple Residual Shrinkage Building Units with LSTM	UCF Crime	97.50	97.10	97.30	97.20
	UCSD	95.70	95.40	95.60	95.50
Multiple Residual Shrinkage Building Units with GRU	UCF Crime	97.80	97.50	97.60	97.55
	UCSD	96.00	95.80	95.90	95.85
Multiple Residual Shrinkage Building Units Gated GRU	**UCF Crime**	**99.95**	**99.95**	**99.95**	**99.80**
	UCSD	**99.98**	**99.98**	**99.98**	**99.98**

Table 16
Ablation study of individual and combined combinations of key components in proposed model.

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Inception v3 + GRU	UCF 50	97.63	97.36	97.14	97.25
	UCF 101	96.88	96.74	96.42	96.58
	HMDB 51	95.13	95.07	94.43	94.75
GRU + Attention	UCF 50	98.38	98.64	98.16	98.40
	UCF 101	97.83	97.52	97.67	97.59
	HMDB 51	96.13	96.02	95.86	95.94
MRSBU + GRU	UCF 50	99.22	99.13	98.97	99.05
	UCF 101	98.94	98.67	98.73	98.70
	HMDB 51	97.83	98.04	97.27	97.65
MRSBU + GRU + Attention	UCF 50	**99.75**	**99.85**	**99.75**	**99.80**
	UCF 101	**99.55**	**99.40**	**99.40**	**99.40**
	HMDB 51	**98.95**	**99.30**	**98.20**	**98.75**

Table 17
Comparative analysis of proposed algorithm using UCF50 dataset.

Methods	Accuracy (%)
Lightweight backbone model for HAR (Varshney & Bakariya, 2022)	98.98
ViT-ReT (Wensel et al., 2023)	97.1
CNN (Ahmad et al., 2023)	99.58
Proposed Multiple Residual Shrinkage Building Units with GRU	**99.75**

Table 18
Comparative analysis of proposed algorithm using UCF101 dataset.

Methods	Accuracy (%)
Multiple CNN Streams (Gangwar et al., 2024)	97.2
Lightweight backbone model for HAR (Varshney & Bakariya, 2022)	98.73
HyEx-Net (Khan et al., 2024)	89.87
Fine-tuned VideoMAE (Gupta & Mahapatra, 2025)	94.15
Multimodal framework (Zhao et al., 2023)	98.35
MVABLSTM (Lim et al., 2023)	99.6
ConvST-LSTM-NET (Jayamohan & Yuvaraj, 2025)	92.8
3DCCA (Praveenkumar et al., 2023)	90.9
Three Stream TDS-BiLSTM (Tan et al., 2022)	94.78
Deep CNN (Yadav et al., 2024)	71.89
ViT-ReT (Wensel et al., 2023)	94.7
FTDS-1DConvNet (Hussain et al., 2024)	88.43
K-NN (Sharma & Singh, 2023)	82.02
Proposed Multiple Residual Shrinkage Building Units with GRU	99.55

Table 19
Comparative analysis of proposed algorithm using HMDB51 dataset.

Methods	Accuracy (%)
Multiple CNN Streams (Gangwar et al., 2024)	85.1
Fusion of DNN (Ullah et al., 2023)	93.7
Lightweight backbone model for HAR (Varshney & Bakariya, 2022)	80.34
HyEx-Net (Khan et al., 2024)	91.77
Fine-tuned VideoMAE (Gupta & Mahapatra, 2025)	78.44
Multimodal framework (Zhao et al., 2023)	83.45
MVABLSTM (Lim et al., 2023)	90.2
ConvST-LSTM-NET (Jayamohan & Yuvaraj, 2025)	91.86
3DCCA (Praveenkumar et al., 2023)	59.5
Three Stream TDS-BiLSTM (Tan et al., 2022)	70.72
Deep CNN (Yadav et al., 2024)	91.79
ViT-ReT (Wensel et al., 2023)	78.4
FTDS-1DConvNet (Hussain et al., 2024)	56.23
K-NN (Sharma & Singh, 2023)	79.98
CNN (Ahmad et al., 2023)	92.70
Proposed Multiple Residual Shrinkage Building Units with GRU	**98.95**

enables deployment on wearable devices, ensuring privacy-preserving monitoring. Many existing HAR algorithms, such as CNN-based methods, struggle because they pass noisy or redundant features through the network and fail to capture long-range temporal dependencies. Hybrid models improve performance but still lack robustness in unconstrained video conditions. By contrast, the proposed model introduces adaptive residual shrinkage, which dynamically suppresses noisy activations, ensuring refined spatial feature representations, while the attention-guided GRU efficiently captures subtle action transitions with lower computational overhead. Theoretical insights into noise reduction and temporal saliency in spatiotemporal learning, combined with practical advantages, provide superior accuracy, efficiency, and robustness across both benchmark and real-world dataset.

The results acquired from the proposed MRSBU-GRU model confirm its superiority over traditional algorithms across multiple benchmark dataset. As presented in Tables 5–7, the proposed model consistently outperforms baseline models such as CNN, RNN, CNN-LSTM, and transformer-based models, achieving the highest performance. These improvements are obtained from adaptive soft-thresholding in MRSBU, which suppresses noisy and redundant features before temporal modeling. Additionally, the integration of GRU with attention allows the model to capture long-range dependencies and dynamically employ importance to relevant features, leading to higher recall and F1-scores. These findings validate that the proposed model improves accuracy and also obtains superior computational efficiency, which is essential for HAR scenarios. As presented in Table 10, the proposed model demonstrates reduced execution time, shorter training time per epoch, and faster inference while consuming less memory compared to other MRSBU-integrated variants. Moreover, Table 11 shows that the proposed model balances trainable parameters and FLOPs per video, supporting its suitability for real-time applications. Generalization ability is further validated through cross-dataset experiments. As presented in Tables 12 and 13, the proposed model consistently maintains high accuracy when trained on UCF101 and tested on UCF50, and vice versa, outperforming traditional models. Evaluations on real-world dataset such as UCF-Crime and UCSD, presented in Table 15, also demonstrate high accuracy, highlighting robustness under unconstrained and noisy environments. Class-wise accuracy and precision results are shown in Figs. 9–26, along with confusion matrices and ROC curves, which shows the model's reliability across different classes and scenarios. Finally, ablation results presented in Table 16 demonstrate that the combination of MRSBU, GRU, and attention yields balanced performance. Evaluations across accuracy, efficiency, and generalization validate the effectiveness of the proposed MRSBU-GRU model. By addressing challenges such as feature redundancy, temporal misalignment, and noisy activations, the proposed model demonstrates both theoretical and practical applicability for human action recognition tasks.

6.1. Error analysis and limitations

Although the proposed MRSBU-GRU model achieves high accuracy across benchmark dataset, certain failure cases are observed where visually similar actions are confused. Figs. 27 and 28 present representative misclassification sequences. These errors generally arise from subtle similarities in hand-face interactions, reliance on background cues, or overlapping motion patterns. This observation shows the challenges of fine-grained activity recognition.

In Fig. 27, the activity HighJump is intermittently misclassified as SkateBoarding, Basketball, or Diving across various frames. This suggests that the proposed model occasionally relied on background features such as sports ground, running tracks, or partial static poses, rather than complete temporal motion, leading to unstable predictions. Although GRU with attention captured temporal dependencies in many cases, subtle frame-to-frame variations introduced noise that caused prediction drift.

In Fig. 28, the reverse error is observed, where Skateboarding is confused with High Jump. The similarities in fast, dynamic body movements and elongated motion trajectories such as jumping or rising actions contribute to this confusion. Additionally, environmental background clutter biased the model, making predictions sensitive to context rather than pure motion cues.

These outcomes highlight three primary challenges of the method:

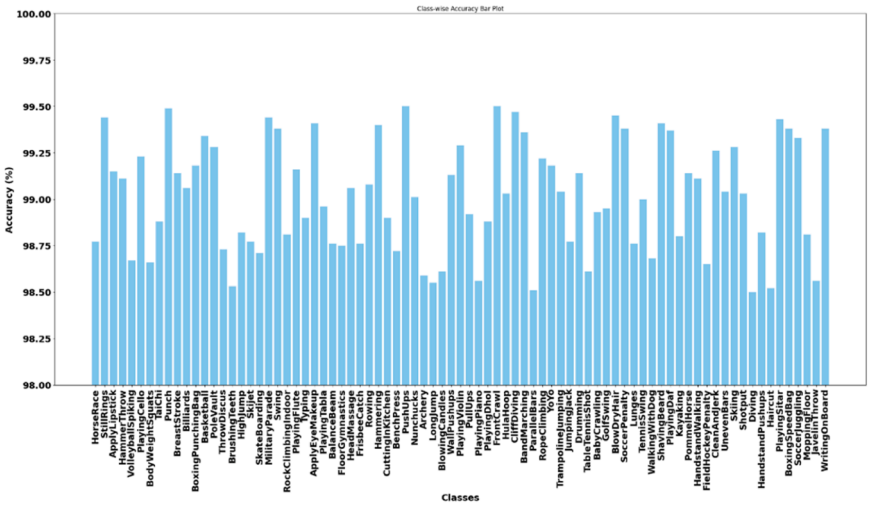

Fig. 15. Class wise accuracy for proposed model across UCF101 dataset action classes.

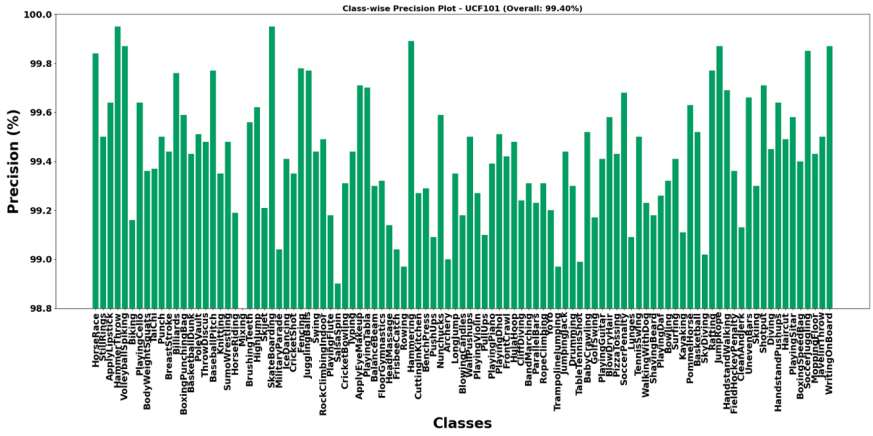

Fig. 16. Class-wise precision of proposed model across UCF101 dataset action classes.

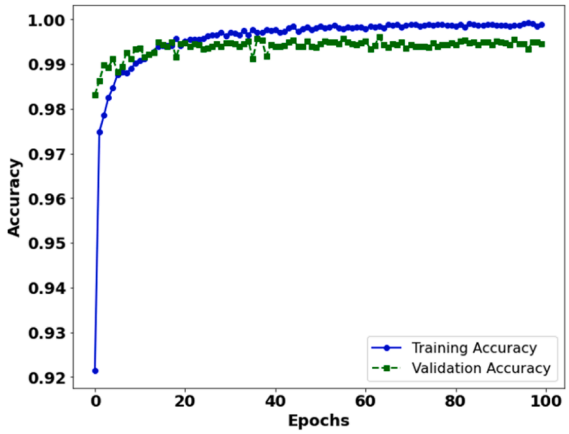

Fig. 17. Accuracy vs Epochs for proposed model across UCF101 dataset.

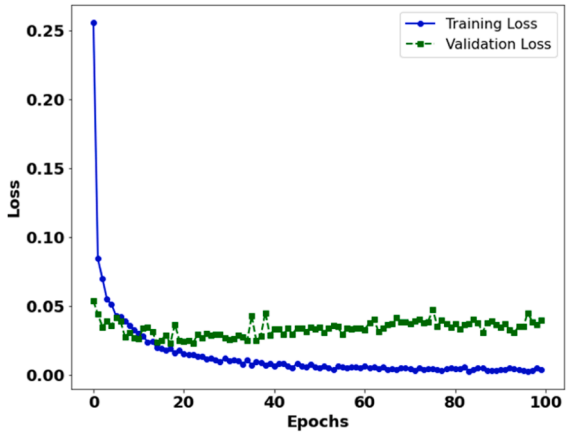

Fig. 18. Loss vs Epochs for proposed model across UCF101 dataset.

Confusion between dynamic sports activities –Actions with similar motion trajectories, such as jumping, running, or flipping, are more prone to misclassification.

Over-reliance on background cues –In several cases, contextual features such as athletic track fields or urban structures strongly influenced classification more than the actual motion sequence.

Fig. 19. Confusion matrix for proposed model across UCF 101 dataset.

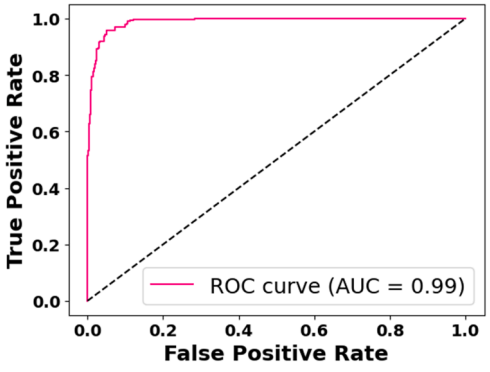

Fig. 20. ROC Curve of proposed model across UCF101 dataset.

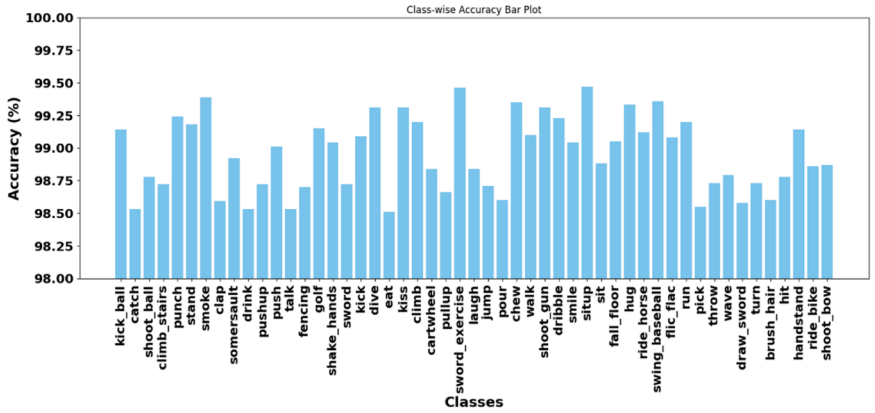

Fig. 21. Class wise accuracy for proposed model across HMDB51 dataset action classes.

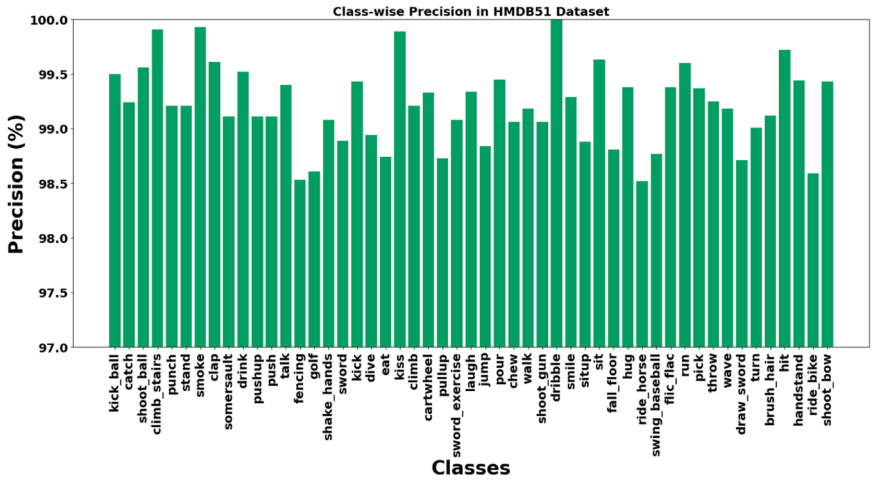

Fig. 22. Class-wise precision of proposed model across HMDB51 dataset action classes.

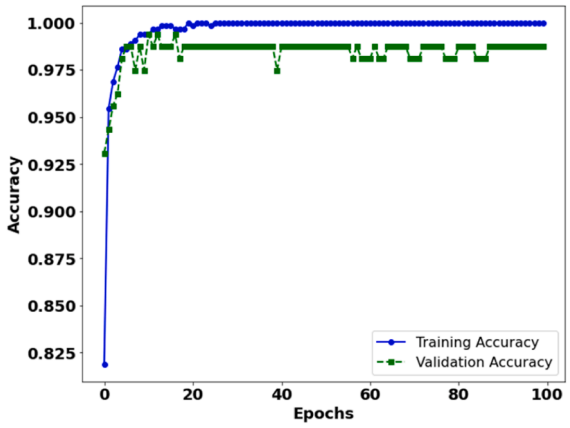

Fig. 23. Accuracy vs Epochs for proposed model across HMDB51 dataset.

Temporal inconsistency across frames –Frame-wise predictions occasionally fluctuated between multiple classes before stabilizing, suggesting that incorporating long-range temporal modeling or optical flow could minimize noise.

Despite these misclassifications, the majority of coarse-grained and visually distinct activities are correctly classified, showing the robustness of the proposed model.

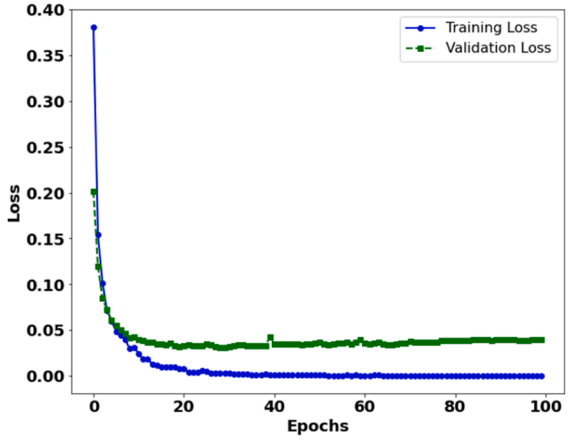

Fig. 24. Loss vs Epochs for proposed model across HMDB51 dataset.

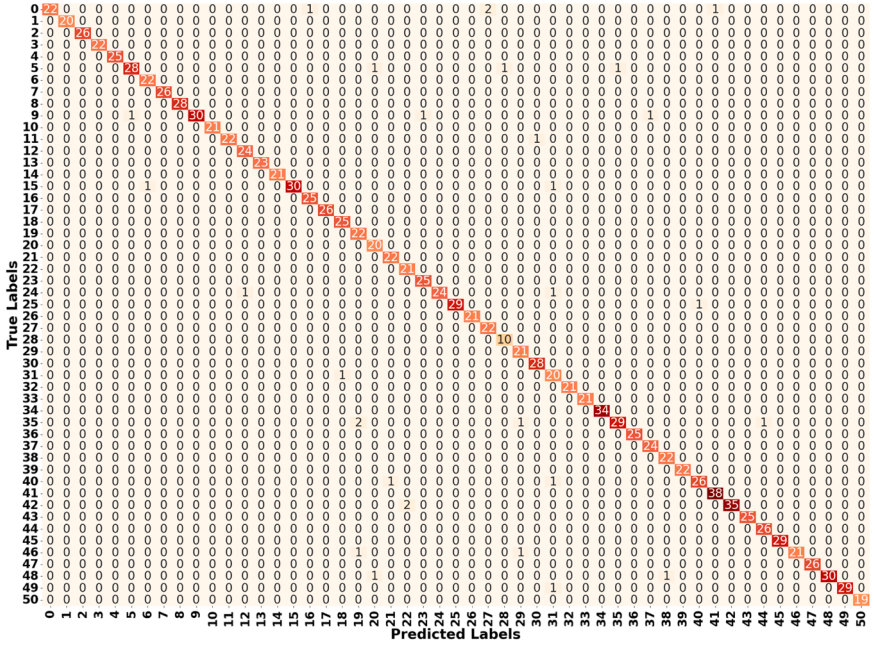

Fig. 25. Confusion matrix for proposed model across HMDB 51 dataset.

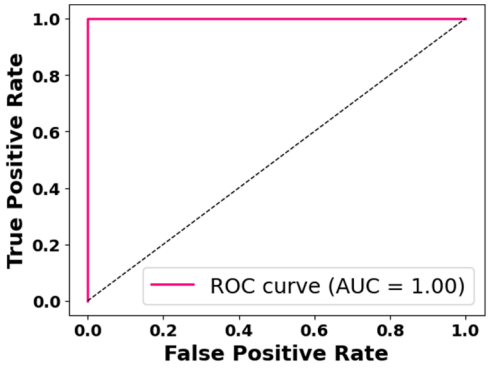

Fig. 26. ROC Curve of proposed model across HMDB51 dataset.

Fig. 27. Frame-wise predictions for HighJump class. Although the ground truth label remains constant, predictions fluctuate between SkateBoarding, Basketball, Diving and HighJump. This represents temporal inconsistency and over-reliance on background cues.

7. Conclusion

Numerous algorithms have been developed for HAR, but these methods often fail to appropriately extract spatial and temporal features to achieve optimal outcomes. In this manuscript, the proposed MRSBU with GRU is implemented for effective recognition of human activities using the UCF50, UCF101, and HMDB51 dataset.

Initially, videos are converted into frames to identify human activities. Features are then extracted using the Inception v3 model, which helps identify human activities within video frames. The RSBU, an improved residual block, learns to suppress less significant features through a soft-thresholding mechanism. It also learns dynamic thresholds to determine whether features are significant. The output of RSBU is flattened and arranged across time steps, and these sequences are fed into a GRU layer that learns temporal

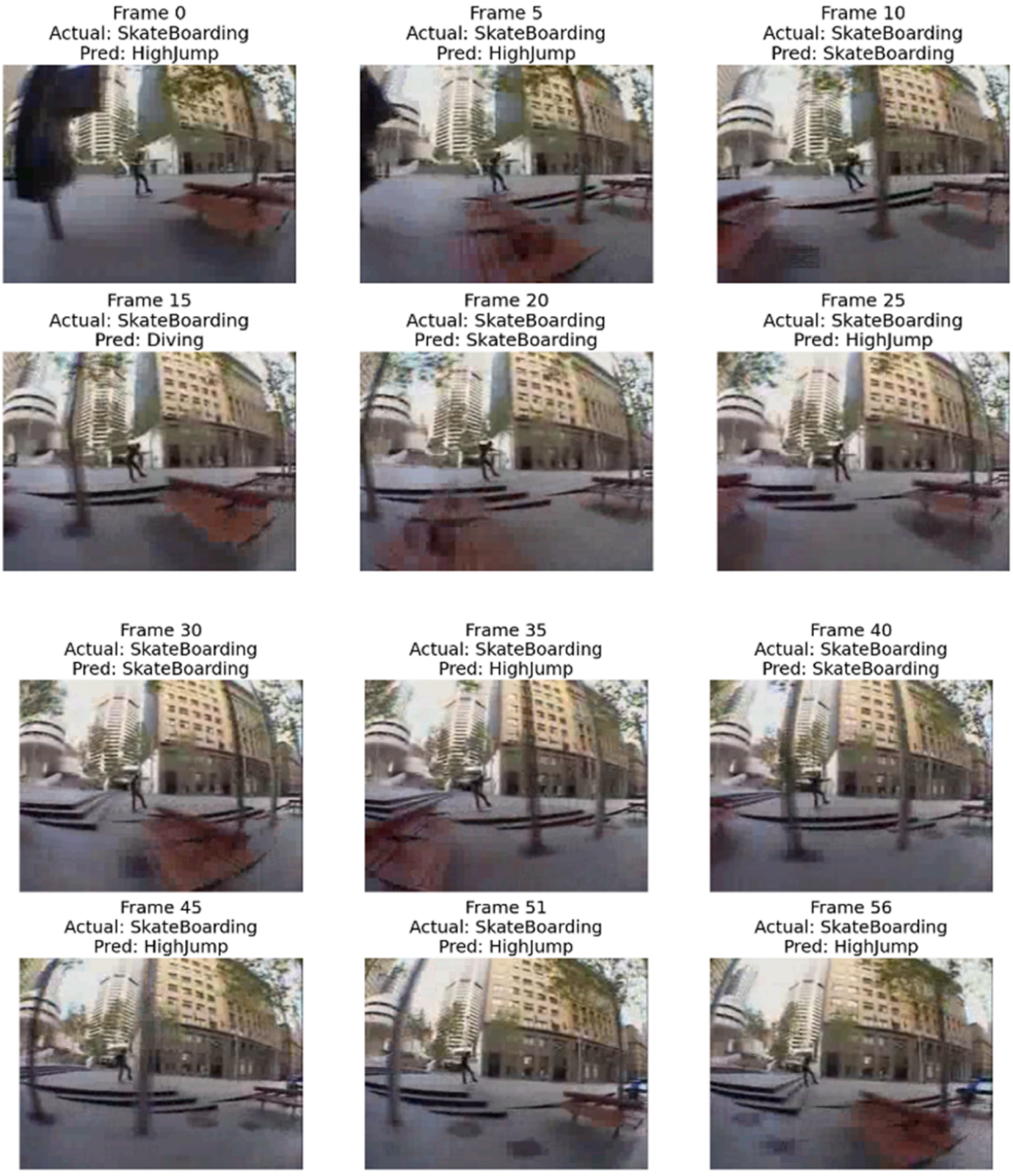

Fig. 28. Frame-wise predictions for SkateBoarding class. The model intermittently predicts HighJump because of similar jumping-like motion patterns and cluttered background features, represents a limitation in fine-grained motion discrimination.

relationships.

7.1. Limitations

Although the proposed MRSBU-GRU model demonstrates superior performance across benchmark and real-world dataset, certain challenges remain. Evaluation is conducted only on publicly available dataset such as UCF50, UCF101, HMDB51, UCF-Crime, and UCSD. While these dataset cover a wide range of activities, they do not capture all possible variations in real-world environments, such as cultural diversity, extreme occlusion, or highly crowded scenes. Although efficiency is validated through execution time, inference time, and memory consumption, deployment on low-power edge devices and resource-constrained IoT system has not been tested.

While the model shows robustness to background noise, its adaptability to entirely unseen environments or multimodal sensor inputs has not been explored.

7.2. Future work

As future work, the developed method will be extended to handle multi-person activity recognition in crowded and occluded scenes. Integration with real-time video processing frameworks will enhance deployment in surveillance and smart environments. Moreover, exploring lightweight architectures for edge devices and incorporating self-supervised learning techniques will improve adaptability and minimize dependence on labeled data.

Declarations

Ethics approval

I/We declare that the work submitted for publication is original, previously unpublished in English or any other language(s), and not under consideration for publication elsewhere.

Consent to participate / Informed consent

Not Applicable.

Consent for publication

I certify that all the authors have approved the paper for release and are in agreement with its content.

Funding

This research received no external funding.

CRediT authorship contribution statement

S. Banushri: Writing – original draft, Software, Resources, Methodology, Data curation, Conceptualization. **R. Jagadeesha:** Writing – original draft, Resources, Project administration, Investigation, Formal analysis.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Appendix A: Basic equations

The fundamental equations utilized in this manuscript are described below. These equations are standard on DL and have moved to appendix for clarity of main text

A.1 ReLU activation

$$f(x) = \max(0, x)$$

ReLU introduces non-linearity through passing positive values when setting negative inputs to zero.

A.2 Pooling Operation

For feature map F, pooling process over region R is described as

Max Pooling

$$P_{max} = \underset{(i,j) \in R}{max} F(i,j)$$

$$P_{avg} = \frac{1}{|R|} \sum_{(i,j) \in R} F(i,j)$$

A.3 Softmax Function

For a given input vector $z = (z_1, z_2, \ldots, z_k)$, softmax probability of class j is

$$\sigma(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

This ensures outputs are normalized to probability values across K classes.

A.4 Cross-Entropy Loss

For a true label y and predicted probability $\hat{y}$, cross-entropy loss is described as:

$$L = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)$$

This calculates divergence among predicted probability distribution and true class labels.

A. 5 GRU Update and Reset Gates

The key operations of GRU are

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

$$\widetilde{h}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h}_t$$

Where the z_t and r_t are update and reset gates and h_t is hidden state at time t.

Data availability

Data will be made available on request.

References

[1] Thakur, D., Biswas, S., Ho, E. S., & Chattopadhyay, S. (2022). Convae-lstm: Convolutional autoencoder long short-term memory network for smartphone-based human activity recognition. *IEEE Access : Practical Innovations, Open Solutions, 10*, 4137–4156. https://doi.org/10.1109/ACCESS.2022.3140373

[2] Paula, L. P. O., Faruqui, N., Mahmud, I., Whaiduzzaman, M., Hawkinson, E. C., & Trivedi, S. (2023). A novel front door security (FDS) algorithm using GoogleNet-BiLSTM hybridization. *IEEE Access : Practical Innovations, Open Solutions, 11*, 19122–19134. https://doi.org/10.1109/ACCESS.2023.3248509

[3] Pareek, G., Nigam, S., & Singh, R. (2024). Modeling transformer architecture with attention layer for human activity recognition. *Neural Computing & Applications, 36*, 5515–5528. https://doi.org/10.1007/s00521-023-09362-7

[4] Lin, H. Y., & Tseng, C. H. (2024). Abnormal activity detection and classification of bus passengers with in-vehicle image sensing. *IEEE Access : Practical Innovations, Open Solutions, 12*, 23057–23065. https://doi.org/10.1109/ACCESS.2024.3365138

[5] Bilal, M., Maqsood, M., Yasmin, S., Hasan, N. U., & Rho, S. (2022). A transfer learning-based efficient spatiotemporal human action recognition framework for long and overlapping action classes. *The Journal of Supercomputing, 78*, 2873–2908. https://doi.org/10.1007/s11227-021-03957-4

[6] Kushwaha, A., Khare, A., & Prakash, O. (2023). Micro-network-based deep convolutional neural network for human activity recognition from realistic and multi-view visual data. *Neural Computing & Applications, 35*, 13321–13341. https://doi.org/10.1007/s00521-023-08440-0

[7] Khaliluzzaman, M., Furquan, M., Zaman Khan, M. S., & Hoque, M. J. (2024). STA-HAR: A spatiotemporal attention-based framework for human activity recognition. *Applied Computational and Intelligence Soft Computing, 2024*, Article 1832298. https://doi.org/10.1155/2024/1832298

[8] Vrskova, R., Hudec, R., & Kamencay, P. (2022a). Sykora Human activity classification using the 3DCNN architecture. *Applied Sciences, 12*, 931. https://doi.org/10.3390/app12020931

[9]M.F. Ahmed, G. He, S. Wang, TriFusion hybrid model for human activity recognition. Signal, image video process. 18 (2024) 8469–8476. https://doi.org/10.1007/s11760-024-03487-5.

[10] Khan, I., Guerrieri, A., Serra, E., & Spezzano, G. (2025). A hybrid deep learning model for UWB radar-based human activity recognition. *Internet Things, 29*, Article 101458. https://doi.org/10.1016/j.iot.2024.101458

[11] Hassan, N., Miah, A. S. M., & Shin, J. (2024). A deep bidirectional LSTM model enhanced by transfer-learning-based feature extraction for dynamic human activity recognition. *Applied Sciences, 14*, 603. https://doi.org/10.3390/app14020603

[12] Dharejo, F. A., Zawish, M., Zhou, Y., Davy, S., Dev, K., Khowaja, S. A., Fu, Y., & Qureshi, N. M. F. (2022). FuzzyAct: A fuzzy-based framework for temporal activity recognition in IoT applications using RNN and 3D-DWT. *IEEE Transanction Fuzzy System, 30*, 4578–4592. https://doi.org/10.1109/TFUZZ.2022.3152106

[13] Aggarwal, S., Bhola, G., & Vishwakarma, D. K. (2024). Weighted voting ensemble of hybrid CNN-LSTM Models for vision-based human activity recognition. *Multimedia Tools Application, 84*, 13255–13293. https://doi.org/10.1007/s11042-024-19582-1

[14] Kumar, R., & Kumar, S. (2023). Multi-view multi-modal approach based on 5s-cnn and bilstm using skeleton, depth and RGB data for human activity recognition. *Wireless Personal Communication, 130*, 1141–1159. https://doi.org/10.1007/s11277-023-10324-4

[15] Pandey, A., & Kumar, P. (2024). Residual deep gated recurrent unit-based attention framework for human activity recognition by exploiting dilated features. *The Visual Computer, 40*, 8693–8712. https://doi.org/10.1007/s00371-024-03266-w

[16] Karim, M., Khalid, S., Aleryani, A., Tairan, N., Ali, Z., & Ali, F. (2024). Hade: Exploiting human action recognition through fine-tuned deep learning methods. *IEEE Access : Practical Innovations, Open Solutions, 12*, 42769–42790. https://doi.org/10.1109/ACCESS.2024.3378515

[17] Liu, H. C., Khairuddin, A. S. M., Chuah, J. H., Zhao, X. M., Wang, X. D., & Fang, L. M. (2024). HCMT: A novel hierarchical cross-modal transformer for recognition of abnormal behavior. *IEEE Access : Practical Innovations, Open Solutions, 12*, 161296–161311. https://doi.org/10.1109/ACCESS.2024.3483896

[18] Yosry, S., Elrefaei, L., ElKamaar, R., & Ziedan, R. R. (2024). Various frameworks for integrating image and video streams for spatiotemporal information learning employing 2D–3D residual networks for human action recognition. *Discover Applied Sciences, 6*, 141. https://doi.org/10.1007/s42452-024-05774-9

[19]N. Jlidi, O. Jemai, T. Bouchrika, Enhancing Human action recognition through transfer learning and body articulation analysis, circuits, systems, and signal processing. (2025) 1–29. https://doi.org/10.1007/s00034-025-03026-8.

[20] Vrskova, R., Hudec, R., Kamencay, P., & Sykora, P. (2022b). A new approach for abnormal human activities recognition based on ConvLSTM architecture. *Sensors, 22*, 2946. https://doi.org/10.3390/s22082946

[21] Bharathi, A., & Sridevi, M. (2023). Human action recognition in complex live videos using graph convolutional network. *Computers and Electrical Engineering, 110*, Article 108844. https://doi.org/10.1016/j.compeleceng.2023.108844

[22] Javed, M. H., Yu, Z., Li, T., Rajeh, T. M., Rafique, F., & Waqar, S. (2022). Hybrid two-stream dynamic CNN for view adaptive human action recognition using ensemble learning. *International Journal of Machine Learning and Cybernetics, 13*, 1157–1166. https://doi.org/10.1007/s13042-021-01441-2
[23] Dastbaravardeh, E., Askarpour, S., Saberi Anari, M., & Rezaee, K. (2024). Channel attention-based approach with autoencoder network for Human action recognition in low-resolution frames. *International Journal of Intelligent Systems, 2024*, Article 1052344. https://doi.org/10.1155/2024/1052344
[24] Zakariah, M., & Alnuaim, A. (2024). Recognizing human activities with the use of convolutional block attention module. *Egyptian Informatics Journal, 27*, Article 100536. https://doi.org/10.1016/j.eij.2024.100536
[25] Amrani, H., Micucci, D., Mobilio, M., & Napoletano, P. (2025). Leveraging dataset integration and continual learning for human activity recognition. *International Journal of Machine Learning and Cybernetics*, 1–22. https://doi.org/10.1007/s13042-025-02569-1
[26] Rani, M., & Kumar, M. (2024). MobileNet for human activity recognition in smart surveillance using transfer learning. *Neural Computing & Applications, 37*, 3907–3924. https://doi.org/10.1007/s00521-024-10882-z
[27] Bhavani, Y. D., & Pagi, V. B. (2025). Human crime activity recognition and shooting weapon detection in video frames using the contour approximation algorithm, and FastDTW classifier. *Cogent Social Sciences, 11*, Article 2465834. https://doi.org/10.1080/23311886.2025.2465834
[28] Gangwar, A., Gonzalez-Castro, V., Alegre, E., Fidalgo, E., & Martinez-Mendoza, A. (2024). DeepHSAR: Semi-supervised fine-grained learning for multi-label human sexual activity recognition. *Information Processing & Management, 61*(5), Article 103800.
[29] Ullah, W., Hussain, T., & Baik, S. W. (2023). Vision transformer attention with multi-reservoir echo state network for anomaly recognition. *Information Processing & Management, 60*(3), Article 103289.
[30] Varshney, N., & Bakariya, B. (2022). Deep convolutional neural model for human activities recognition in a sequence of video by combining multiple CNN streams. *Multimedia Tools Application, 81*, 42117–42129. https://doi.org/10.1007/s11042-021-11220-4
[31] Khan, M. A., Javed, K., Khan, S. A., Saba, T., Habib, U., Khan, J. A., & Abbasi, A. A. (2024). Human action recognition using fusion of multiview and deep features: An application to video surveillance. *Multimedia Tools Application, 83*, 14885–14911. https://doi.org/10.1007/s11042-020-08806-9
[32] Gupta, S., & Mahapatra, R. P. (2025). Hyex-net: Hybrid efficient Xception network for human activity recognition using multimodalities. *Cluster Computing, 28*, 240. https://doi.org/10.1007/s10586-024-04998-z
[33] Zhao, H., Liu, J., & Wang, W. (2023). Research on human behavior recognition in video based on 3DCCA. *Multimedia Tools Application, 82*, 20251–20268. https://doi.org/10.1007/s11042-023-14355-8
[34] Lim, K. M., Lee, C. P., Tan, K. S., Alqahtani, A., & Ali, M. (2023). Fine-tuned temporal dense sampling with 1d convolutional neural network for human action recognition. *Sensors, 23*, 5276. https://doi.org/10.3390/s23115276
[35] Jayamohan, M., & Yuvaraj, S. (2025). A novel human action recognition using Grad-CAM visualization with gated recurrent units. *Neural Computing & Applications*, 1–16. https://doi.org/10.1007/s00521-025-10978-0
[36] Praveenkumar, S. M., Patil, P., & Hiremath, P. S. (2023). A novel algorithm for human action recognition in compressed domain using attention-guided approach. *Journal of the Real-Time Image Processing, 20*, 122. https://doi.org/10.1007/s11554-023-01374-9
[37] Tan, K. S., Lim, K. M., Lee, C. P., & Kwek, L. C. (2022). Bidirectional long short-term memory with temporal dense sampling for human action recognition. *Expert System Application, 210*, Article 118484. https://doi.org/10.1016/j.eswa.2022.118484
[38] Yadav, R. K., Daniel, A., & Semwal, V. B. (2024). Enhancing Human activity detection and classification using fine tuned attention-based transformer models. *SN Computing Science, 5*, 1–21. https://doi.org/10.1007/s42979-024-03445-5
[39] Wensel, J., Ullah, H., & Munir, A. (2023). Vit-ret: Vision and recurrent transformer neural networks for human activity recognition in videos. *IEEE Access : Practical Innovations, Open Solutions, 11*, 72227–72249. https://doi.org/10.1109/ACCESS.2023.3293813
[40] Hussain, A., Khan, S. U., Khan, N., Shabaz, M., & Baik, S. W. (2024). AI-driven behavior biometrics framework for robust human activity recognition in surveillance systems. *Engineering Applications of Artificial Intelligence, 127*, Article 107218. https://doi.org/10.1016/j.engappai.2023.107218
[41] Sharma, A., & Singh, R. (2023). ConvST-LSTM-Net: Convolutional spatiotemporal LSTM networks for skeleton-based human action recognition. *International Journal of Multimedia Information Retrieval, 12*, 34. https://doi.org/10.1007/s13735-023-00301-9
[42] Ahmad, T., Wu, J., Alwageed, H. S., Khan, F., Khan, J., & Lee, Y. (2023). Human activity recognition based on deep-temporal learning using convolution neural networks features and bidirectional gated recurrent unit with features selection. *IEEE Access, 11*, 33148–33159. https://doi.org/10.1109/ACCESS.2023.3263155
[43] Göçmen, O., & Akata, M. E. (2023). Polygonized silhouettes and polygon coding based feature representation for Human action recognition. *IEEE Access : Practical Innovations, Open Solutions, 11*, 57021–57036. https://doi.org/10.1109/ACCESS.2023.3283458
[44] Uddin, M. A., Talukder, M. A., Uzzaman, M. S., Debnath, C., Chanda, M., Paul, S., Islam, M. M., Khraisat, A., & Alazab S. Aryal, A. (2024). Deep learning-based human activity recognition using CNN, ConvLSTM, and LRCN. *International Journal of Cognitive in Computer Engineering, 5*, 259–268. tps://doi.org/10.1016/j.ijcce.2024.06.004.
[45]UCF101 dataset 2025:Https://www.kaggle.com/dataset/matthewjansen/ucf101-action-recognition (Accessed on April 2025).
[46]UCF50 dataset 2025: Https://www.kaggle.com/dataset/vineethakkinapalli/ucf50-action-recognition-dataset (Accessed on April 2025).
[47]HMDB51 dataset 2025:Https://www.kaggle.com/dataset/easonlll/hmdb51 (Accessed on April 2025).
[48] Ramaneswaran, S., Srinivasan, K., Vincent, P. D. R., & Chang, C. Y. (2021). Hybrid inception v3 XGBoost model for acute lymphoblastic leukemia classification. *Computational and Mathematical Methods in Medicine, 2021*, Article 2577375. https://doi.org/10.1155/2021/2577375