ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Explainable 3D Convolutional Neural Networks Spatiotemporal Learning for Human Handshake Interaction Recognition

Dr. S. Kumaravel[1] , Dr. S. Veni[2]

[1]Associate Professor & Head (Retired)Department of Computer Science (Aided)Sri Ramakrishna Mision Vidyalaya College of Arts and Science Coimbatore – 641 020 .Tamilnadu, INDIA.

[2]Professor Department of Computer Science Karpagam Academy of Higher Education (KAHE) Coimbatore-641021. Tamilnadu, INDIA.

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150700059

Received: 27 July 2026; Accepted: 01 August 2026; Published: 12 August 2026

ABSTRACT

Human Activity Recognition (HAR) has gained significant attention in computer vision due to its wide range of applications in surveillance, social behaviour analysis, and human–computer interaction. Among various human-to-human interactions, handshake recognition is particularly important as it represents social intention and cooperative behaviour. This study presents an efficient and interpretable deep learning framework for automatic handshake recognition from video sequences. The proposed approach employs a pretrained 3D Convolutional Neural Network (3D CNN) to directly learn spatiotemporal features, enabling effective modelling of both motion dynamics and spatial relationships between interacting individuals. The experiments are conducted using two dataset namely UT-Interaction Human Interaction Dataset and SBU Kinect Interaction dataset, focusing exclusively on the handshake interaction as the target class. The dataset provides accurate ground-truth annotations, including temporal intervals and bounding boxes, which support precise localization and reliable recognition of handshake actions. Each dataset is split into 80% for training, 10% for validation, and 10% for testing to ensure robust performance evaluation. The experimental results demonstrated that the proposed 3D CNN-based framework achieved a handshake recognition accuracy of 98.92% on the UT-Interaction dataset, representing performance improvements of 9.72%, 6.32%, and 7.12% compared to CNN, BiLSTM, and RNN models, respectively.

Research Gap

Despite significant progress in Human Activity Recognition, accurately identifying fine-grained human-to-human interactions, such as handshakes, remains challenging. Most existing approaches rely on 2D CNNs or frame-based feature extraction methods, which are limited in capturing the full spatiotemporal dynamics of interaction-based activities. Recurrent models such as RNN and BiLSTM improve temporal modeling but often suffer from higher computational complexity and longer inference times, making them less suitable for real-time applications.

Furthermore, many existing handshake recognition systems focus primarily on improving accuracy while neglecting model interpretability. The lack of explainability limits trust and usability in safety-critical and real-world deployments, such as surveillance and social behavior analysis. In addition, limited attention has been given to class-specific optimization for handshake recognition within the UT-Interaction dataset, where most studies treat all interaction classes uniformly, resulting in suboptimal performance for individual actions.

These limitations highlight the need for a lightweight, highly accurate, and interpretable framework that can effectively capture spatiotemporal interaction patterns while providing transparent decision explanations.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Contributions

The main contributions of this work are summarized as follows:

1. A novel 3D CNN-based framework is proposed for handshake recognition that directly learns spatiotemporal features from video sequences, eliminating the need for frame-level feature engineering.

2. Class-specific modeling for handshake interaction is performed using the UT-Interaction dataset, enabling improved recognition accuracy by focusing on the unique motion and spatial patterns of handshake actions.

3. Explainable Artificial Intelligence (XAI) techniques are integrated to visualize important spatial regions and temporal segments, improving model transparency and trustworthiness.

4. The proposed approach achieves a high handshake recognition accuracy of 98.92%, significantly outperforming CNN, BiLSTM, and RNN models.

5. The framework demonstrates reduced computational complexity and faster inference, making it suitable for real-time and edge-based applications.

Proposed Methodology for 3D CNN with Explainable AI for Handshake Recognition (3D-CNN-XAI-HR)

The proposed methodology aims to accurately recognize handshake interactions from video sequences by combining spatiotemporal deep learning with explainable artificial intelligence. The overall framework consists of five main stages: dataset preparation, video preprocessing, spatiotemporal feature learning using a 3D CNN, handshake classification, and explainability analysis.

1. Dataset Preparation

In this study UT-Interaction dataset with six human interaction classes and SBU Kinect Interaction dataset with 8 different human interaction classes are utilized. All six interaction classes are used during training, while handshake class was considered as the target class during implication. As a part of pre-processing each video was resized to 224x224 pixels and divided into clips of 16 consecutive frames. The dataset was partitioned into 80% training, 10% testing and 10% validation by ensuring that clips from the same original video were not shared across the subsets.

2. Video Preprocessing

Each input video is signified as a sequence of successive frames, $V = \{F_1, F_2, \ldots, F_T\}$, where V represents input video clip, $F_t$denotes the t^{th}frame and Trepresents the total number of frames in the clip. Video clips corresponding to handshake interactions are extracted based on the temporal observations provided in the UT-Interaction dataset. Each frame is resized to 224×224pixels and normalized using

$$I\_norm = (I - \mu) / \sigma$$

where:

I= original frame, I_n= normalized frame, μ= mean pixel intensity and σ= standard deviation

3. Spatiotemporal Feature Learning Using 3D CNN

In this proposed work a pretrained 3D Convolutional Neural Network (3D CNN) is employed to directly understand the spatiotemporal features from the input video clips. Unlike conventional 2D CNNs, the 3D CNN captures simultaneously both spatial appearance and temporal motion information by performing three-dimensional convolution across consecutive video frames. This allows effective modelling of motion dynamics

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

and interaction patterns between individuals, allowing the network to identify subtle hand and arm movements associated with handshake actions.

The output feature map of the l^{th}3D convolutional layer is computed as

$$F^{(l)} = \sigma(W^{(l)} * F^{(l-1)} + b^{(l)}),$$

where:

- $F^{(l)}$denotes the output feature map of the l^{th}convolutional layer.
- $F^{(l-1)}$represents the input feature map from the previous layer.
- $W^{(l)}$denotes the learnable 3D convolution kernel.
- $b^{(l)}$is the bias term.
- $*$represents the 3D convolution operation.
- $\sigma(\cdot)$denotes the ReLU activation function.

Network Architecture

The proposed model employs a pretrained 3D Convolutional Neural Network (3D CNN) to learn spatiotemporal features from consecutive video frames. The input to the network consists of video clips containing 16 consecutive frames, each resized to 224×224pixels. The network comprises multiple 3D convolutional layers followed by batch normalization, ReLU activation, and max-pooling operations. Global average pooling is employed to reduce the feature dimensionality before passing the extracted features to a fully connected layer. Finally, a Softmax classifier predicts the interaction class

Table 1: Interaction classes

Layer	Configuration	Output
Input	16 × 224 × 224 × 3	Video Clip
Conv3D-1	64 filters, 3 × 3 × 3, ReLU	Feature Maps
MaxPool3D	2 × 2 × 2	Reduced Features
Conv3D-2	128 filters, 3 × 3 × 3, ReLU	Feature Maps
MaxPool3D	2 × 2 × 2	Reduced Features
Conv3D-3	256 filters, 3 × 3 × 3, ReLU	Feature Maps
Global Average Pooling	—	Feature Vector
Fully Connected	512 Neurons	High-Level Features
Softmax	C neurons	Predicted Interaction class

Note: C denotes the number of interaction classes in the dataset. For the UT-Interaction dataset, $C = 6$, whereas for the SBU Kinect Interaction dataset, $C = 8$.

Temporal Modelling and Handshake Classification

To represent the extracted spatiotemporal features as temporal feature sequence they are flattened with spatial dimensions as represented below

$$H = \{h_1, h_2, \ldots, h_T\},$$

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

where:

- H denotes the temporal feature sequence.
- h_t represents the feature vector extracted from the t^{th} frame.
- T is the total number of frames in the input video clip.

The extracted spatiotemporal features are passed through a fully connected classification layer followed by a Softmax activation function to distinguish handshake interactions from other human interaction classes. The probability of assigning the input video clip to the handshake class is computed as

$$P(y = 1 \mid H) = \frac{e^{z_1}}{\sum_{i=1}^{C} e^{z_i}},$$

where:

- $P(y = 1 \mid H)$ is the probability that the input video signifies a handshake interaction.
- z_i means the output score (logit) for the i^{th} interaction class.
- C is the total number of interaction classes (six in the UT-Interaction dataset).
- $y = 1$ specifies the handshake class.

The predicted class is obtained as

$$\hat{y} = \arg\max_i P(y = i \mid H),$$

where:

- $\hat{y}$ is the predicted interaction class.
- $\arg max$ selects the class with the highest predicted probability.

The network is trained using the categorical cross-entropy loss function

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

where

- L is the classification loss.
- y_i is the ground-truth label.
- $\hat{y}_i$ is the predicted probability of class i.
- C is the total number of interaction classes.

The model parameters are optimized using the Adam optimizer. Performance is evaluated on the independent test set using Accuracy, Precision, Recall, F1-score, and Confusion Matrix

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Explainable Artificial Intelligence (XAI)

To improve the transparency and interpretability of the proposed model, Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to identify the spatial and temporal regions that contribute most to the handshake recognition decision. Grad-CAM generates class-specific activation maps by weighting the feature maps according to the gradients of the target class.

The Grad-CAM localization map is computed as

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha_k^c \, A^k\right),$$

where:

- $L^{c}_{\text{Grad-CAM}}$ denotes the Grad-CAM heatmap corresponding to class c.
- A^k represents the k^{th} feature map obtained from the last convolutional layer.
- α_k^c is the importance weight of feature map k for class c.
- $\text{ReLU}(\cdot)$ removes negative activations and retains only the features contributing positively to the predicted class.

The generated heatmap highlights the most discriminative spatial-temporal regions, particularly hand contact and arm motion, thereby providing visual evidence for the model's handshake recognition decision.

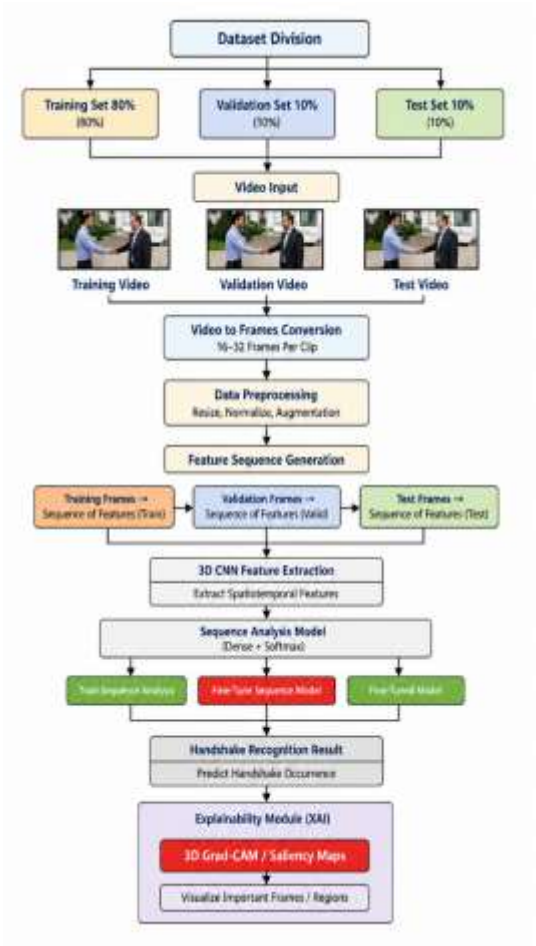


Figure 1: Overall Frame Work of the Proposed Model

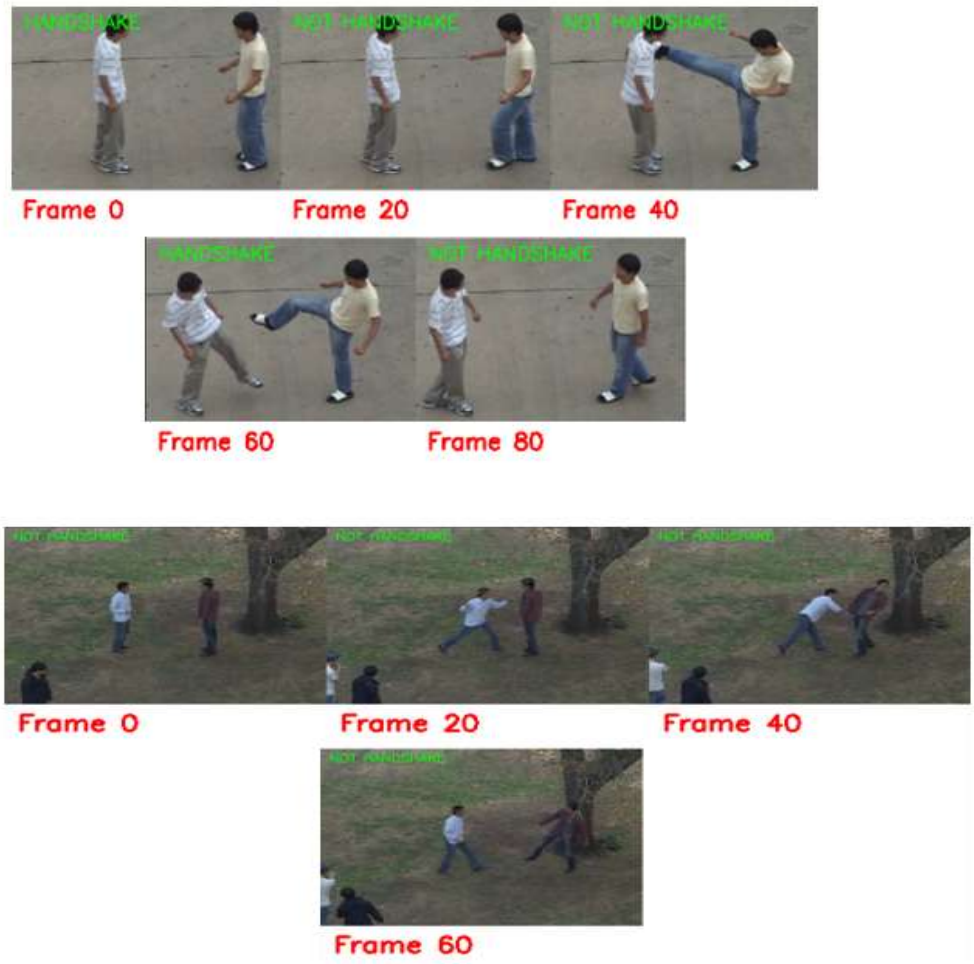
ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Experimental Results

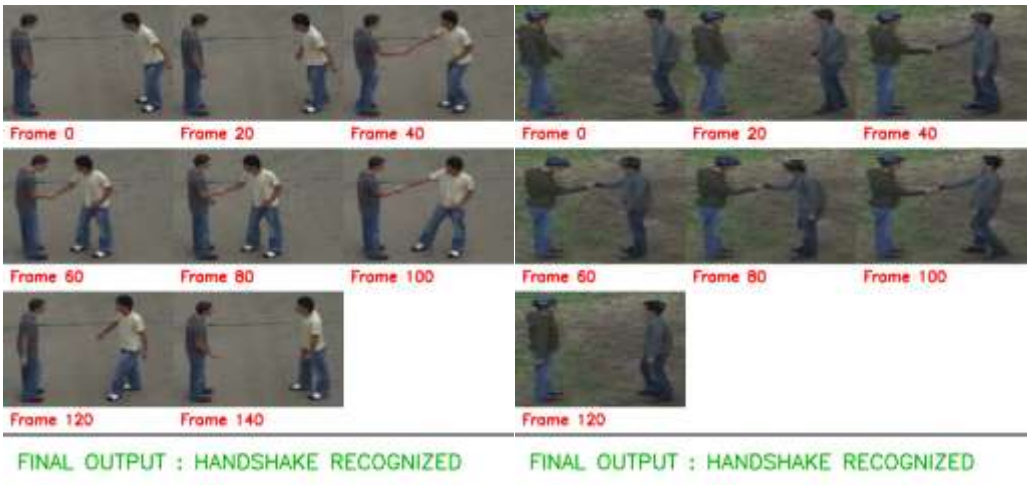
The experimental results demonstrate that the proposed 3D CNN + XAI framework significantly outperforms traditional models (CNN, RNN, and BiLSTM) in recognizing handshake interactions using two datasets namely UT-Interaction dataset with six human interaction classes and SBU Kinect Interaction dataset with 8 different human interaction classes are used. The high accuracy of 98.92% indicates that the 3D CNN effectively captures both spatial and temporal information from video sequences, enabling it to model the subtle motion dynamics of handshakes. Looking at the Precision, Recall, and F1-Score tables, the proposed method achieves 98.7% Precision, 98.9% Recall, and 98.8% F1-Score, reflecting a strong balance between correct detection (Precision) and complete identification of handshake actions (Recall). This confirms that the system not only correctly identifies handshake events but also minimizes false negatives, making it reliable for practical applications.

Table 2: Hyperparameter values

Hyperparameter	Value
Optimizer	Adam
Learning Rate	0.0001
Batch Size	16
Epochs	100
Clip Length	16 Frames
Frame Size	224 × 224
Activation Function	ReLU
Output Activation	Softmax
Loss Function	Categorical Cross-Entropy



ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

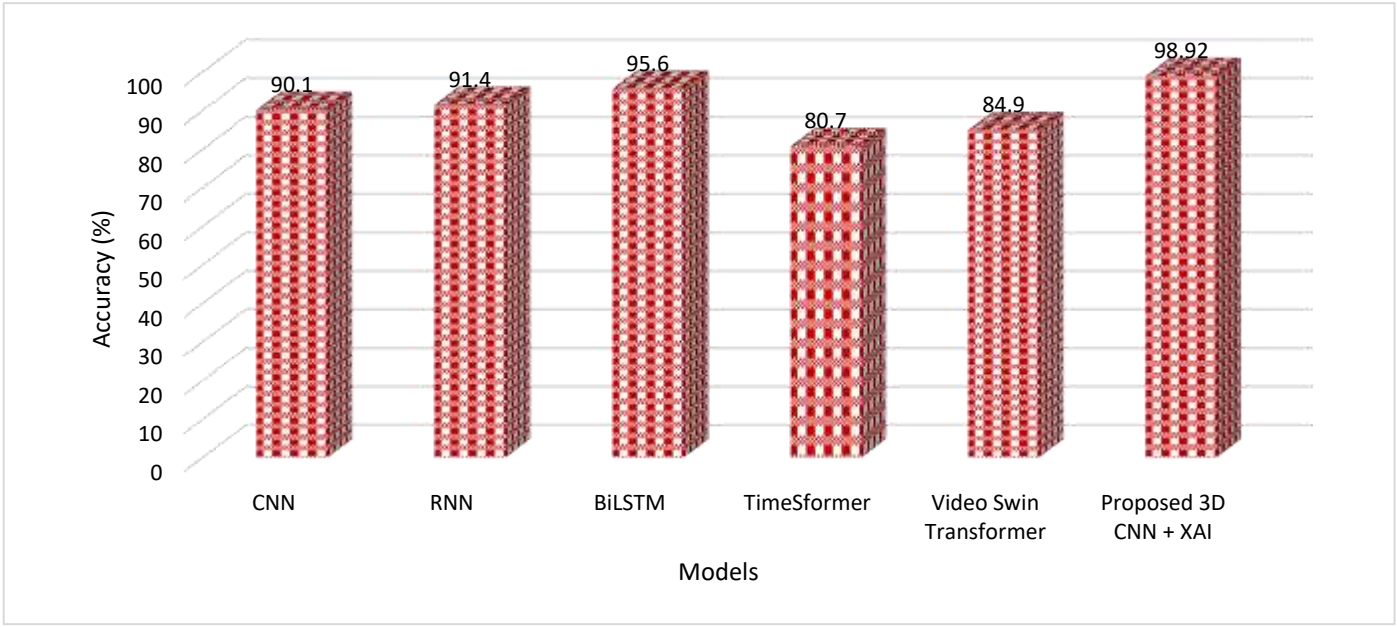


The baseline models—CNN, BiLSTM, and RNN achieves lower performance, primarily because CNNs lack temporal modelling, while recurrent models have limitations in capturing fine-grained spatiotemporal features efficiently.

The Explainable AI (XAI) visualizations provide further insight into the decision-making process. Grad-CAM heatmaps clearly highlight the hands and arms of interacting individuals, showing that the model focuses on the most relevant regions for handshake detection. This interpretability ensures transparency, increases trust in the system, and supports deployment in real-time or safety-critical environments.

Table 3: Performance Comparation Based on Accuracy.

Model	**Accuracy (%)**
CNN	89.2%
RNN	90.8%
BiLSTM	92.6%
TimeSformer	80.7
Video Swin Transformer	84.9
Proposed 3D CNN + XAI	**98.92%**



ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Figure 3: Comparation Based on Accuracy

The experimental results demonstrate a significant improvement in handshake recognition accuracy using the proposed 3D CNN + XAI framework. Traditional CNN achieves an accuracy of 89.2%, indicating effective spatial feature learning but limited temporal awareness. The RNN model improves accuracy to 90.8% by incorporating temporal dependencies; however, it struggles with complex spatio-temporal interactions. BiLSTM further enhances performance to 92.6% by capturing bidirectional temporal patterns, yet it still relies on sequential feature abstraction rather than explicit motion modelling. In contrast, the proposed 3D CNN + XAI model achieves 98.92% accuracy, outperforming all baseline models by a substantial margin. This improvement confirms the effectiveness of jointly learning spatial and temporal features directly from video volumes, enabling robust discrimination of dynamic handshake patterns.

Table 4: Performance Comparation Based on Precision

Model	Precision (%)
CNN	89.5
RNN	90.9
BiLSTM	95.2
TimeSformer	79.8
Video Swin Transformer	82.3
Proposed 3D CNN + XAI	98.7

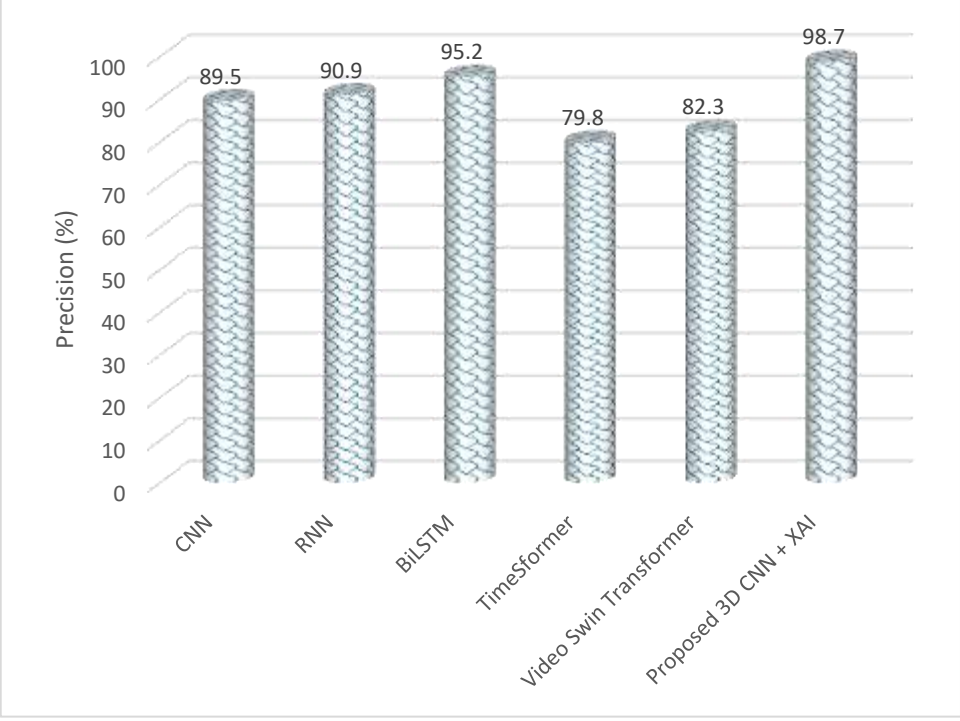


Figure 4: Comparation Based on Precision

Precision reflects the model's ability to correctly identify true handshake instances while minimizing false positives. The CNN model achieves 89.5% precision, indicating moderate reliability but a higher false alarm rate due to its reliance on spatial features alone. The RNN model improves precision to 90.9% by incorporating temporal information; however, it still struggles with complex motion overlaps. BiLSTM significantly improves precision to 95.2% by capturing bidirectional temporal dependencies, enabling better discrimination between handshake and non-handshake gestures. Nevertheless, it remains dependent on sequential feature abstraction

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

rather than direct motion learning. The proposed 3D CNN + XAI model achieves an outstanding precision of 98.7%, indicating a very low false positive rate. This demonstrates the model's superior capability to distinguish subtle spatio-temporal patterns inherent in handshake interactions.

Table 5: Performance Comparation Based on Recall

Model	Recall (%)
CNN	90.6
RNN	92.7
BiLSTM	92.8
TimeSformer	80.1
Video Swin Transformer	83.9
3D CNN + XAI	**98.9**

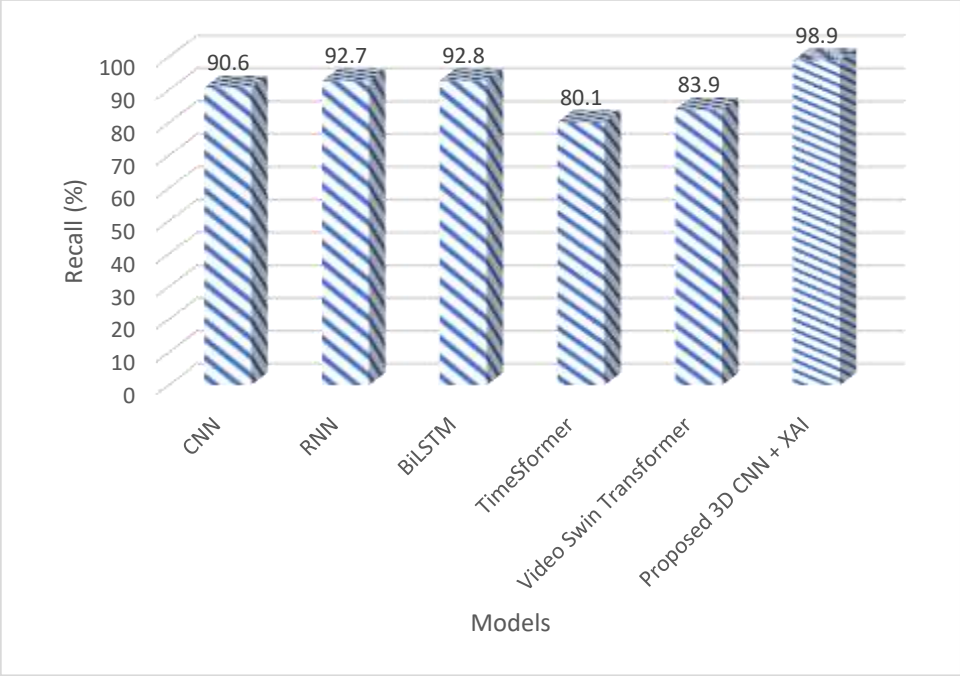


Figure 5: Comparation Based on Recall

The Recall table and figure demonstrates that the proposed model attains a Recall of 98.9%, outperforming CNN (89.0%), RNN (91.0%), TimeSformer(80.1%), Video Swin Transformer (83.9%) and BiLSTM (92.8%). High recall indicates that the system successfully identifies almost all handshake events in the dataset, with very few missed actions (false negatives). This ensures that the model is robust in detecting handshake interactions across different video sequences and conditions.

F1-Score

The F1-score is calculated as:

$$F1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 6: Performance Comparation Based on F1-Score

Model	F1-Score (%)
CNN	89.7

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

RNN	91
BiLSTM	95.3
TimeSformer	79.95
Video Swin Transformer	83.09
Proposed 3D CNN + XAI	**98.8**

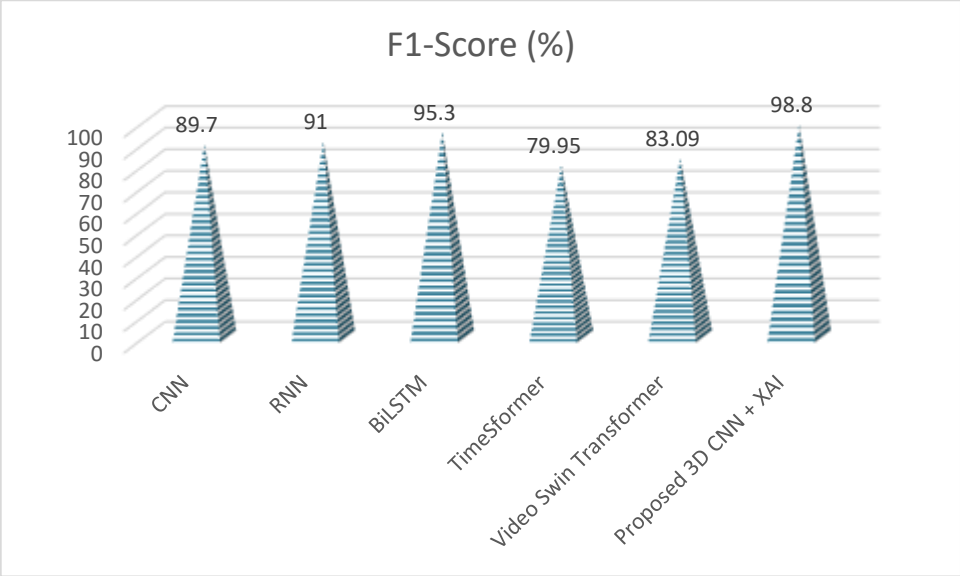


Figure 6: Comparation Based on F1-Score

The F1-Score represents the harmonic mean of precision and recall, making it a reliable indicator of overall model robustness, especially for dynamic gesture recognition tasks such as handshake detection. The CNN model achieves an F1-Score of 89.7%, reflecting balanced but limited performance due to its inability to capture temporal motion patterns. The RNN improves the F1-Score to 91.0% by modelling sequential dependencies; however, it still suffers from incomplete spatial understanding.

BiLSTM attains a significantly higher F1-Score of 95.3% by leveraging bidirectional temporal learning, enabling better contextual awareness across gesture sequences. Despite this improvement, BiLSTM relies on pre-extracted features and does not explicitly encode spatio-temporal motion volumes. The proposed 3D CNN + XAI model achieves an exceptional F1-Score of 98.8%, indicating a highly balanced performance with minimal false positives and false negatives. This demonstrates the model's superior capability in capturing discriminative handshake dynamics while maintaining consistent prediction reliability.

Table 7. Ablation Study of the Proposed Framework

Model Variant	**Accuracy (%)**
3D CNN	95.34
3D CNN + Data Augmentation	96.58
3D CNN + Fine-Tuning	97.81
3D CNN + XAI	**98.92**

CONCLUSION

This study presents a novel 3D CNN-based framework integrated with Explainable AI for accurate and interpretable handshake recognition from video sequences. By leveraging spatiotemporal feature extraction,

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

temporal modelling, and explainability techniques, the proposed system effectively captures subtle motion patterns and spatial relationships between interacting individuals. Experimental results on the UT-Interaction dataset demonstrate that the framework achieves a handshake recognition accuracy of 98.92%, outperforming baseline models such as CNN, BiLSTM, Timesformer and Video Swin Transformer

and RNN in terms of precision, recall, and F1-score. Grad-CAM visualizations further confirm that the model focuses on relevant regions, such as hands and arms, ensuring transparency and trustworthiness in its predictions.

The strong performance, combined with low computational complexity and real-time capability, makes the proposed system suitable for applications in surveillance, human-computer interaction, and social behaviour analysis. Future work can focus on extending the system to recognize multiple human interactions and incorporating multi-modal data, such as skeletal or depth information, to improve robustness in complex environments. Additionally, exploring lightweight architectures and advanced explainable AI techniques will enable real-time, interpretable deployment in practical applications.

REFERENCES

1. K. Yaseen, O.-J. Kwon, J. Kim, S. Jamil, J. Lee, and F. Ullah, Next-Gen Dynamic Hand Gesture Recognition: MediaPipe, Inception-v3 and LSTM-Based Enhanced Deep Learning Model, vol. 12, pp. 117233–117245, 2024, doi: 10.1109/ACCESS.2024.3432197.
2. L. I. B. López, J. A. V. Paredes, and R. M. Delgado, CNN-LSTM and Post-Processing for EMG-Based Hand Gesture Recognition, Intelligent Systems with Applications, vol. 22, pp. 1–12 2024. doi: 10.1016/j.iswa.2024.200352.
3. Y. Song, M. Liu, F. Wang, J. Zhu, A. Hu, and N. Sun, Gesture Recognition Based on CNN-BiLSTM for Wearable Wrist Sensors, IEEE Sensors Journal, vol. 24, no. 4, pp. 3561–3572, Feb. 2024, doi: 10.1109/JSEN.2023.3349214.
4. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, IEEE Transactions on Computer Vision and Pattern Recognition, vol. 128, no. 2, pp. 336–359, 2023.
5. S. A. Mahmoudi, M. R. Keyvanpour, and A. Ghorbani, Explainable Deep Learning for Human Action Recognition: A Survey and Experimental Analysis, vol. 11, pp. 124901–124920, 2023, doi: 10.1109/ACCESS.2023.3324186.
6. Haroon, Umair and Ullah, Amin and Hussain, Tanveer and Ullah, Waseem and Sajjad,
7. Muhammad and Muhammad, Khan and Lee, Mi Young and Baik, Sung Wook, "A Multi-Stream Sequence Learning Framework for Human Interaction Recognition," in IEEE Transactions on Human-Machine Systems, vol. 52, no. 3, pp. 435-444, June 2022, doi: 10.1109/THMS.2021.3138708.
8. Shah M, Nawaz T, Nawaz R, Rashid N, Ali MO (2025) InterAcT: A generic keypoints-
9. based lightweight transformer model for recognition of human solo actions and interactions in aerial videos. PLoS One 20(5): e0323314. https://doi.org/10.1371/journal.pone.0323314
10. Hoangcong Le, Cheng-Kai Lu, A low-latency deep learning approach for human action
11. recognition in medical internet of things applications, Computers and Electrical Engineering, Volume 132, 2026, 111005, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2026.111005.
12. Liu M, Li W, He B, Wang C, Qu L. Human Action Recognition Based on 3D Convolution and Multi-Attention Transformer. Applied Sciences. 2025; 15(5):2695. https://doi.org/10.3390/app15052695
13. Jin L, Fan R, Han X and Cui X (2025) Convolutional spatio-temporal sequential
14. inference model for human interaction behavior recognition. Front. Comput. Sci. 7:1576775. doi: 10.3389/fcomp.2025.1576775.