Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.056767

ARTICLE

# Human Interaction Recognition in Surveillance Videos Using Hybrid Deep Learning and Machine Learning Models

**Vesal Khean[1], Chomyong Kim[2], Sunjoo Ryu[2], Awais Khan[1], Min Kyung Hong[3], Eun Young Kim[4], Joungmin Kim[5] and Yunyoung Nam[3,*]**

[1]Department of ICT Convergence, Soonchunhyang University, Asan, 31538, Republic of Korea

[2]ICT Convergence Research Center, Soonchunhyang University, Asan, 31538, Republic of Korea

[3]Emotional and Intelligent Child Care Convergence Center, Soonchunhyang University, Asan, 31538, Republic of Korea

[4]Department of Occupational Therapy, Soonchunhyang University, Asan, 31538, Republic of Korea

[5]College of Hyangsul Nanum, Soonchunhyang University, Asan, 31538, Republic of Korea

*Corresponding Author: Yunyoung Nam. Email: ynam@sch.ac.kr

Received: 29 July 2024 Accepted: 29 August 2024 Published: 15 October 2024

## ABSTRACT

Human Interaction Recognition (HIR) was one of the challenging issues in computer vision research due to the involvement of multiple individuals and their mutual interactions within video frames generated from their movements. HIR requires more sophisticated analysis than Human Action Recognition (HAR) since HAR focuses solely on individual activities like walking or running, while HIR involves the interactions between people. This research aims to develop a robust system for recognizing five common human interactions, such as hugging, kicking, pushing, pointing, and no interaction, from video sequences using multiple cameras. In this study, a hybrid Deep Learning (DL) and Machine Learning (ML) model was employed to improve classification accuracy and generalizability. The dataset was collected in an indoor environment with four-channel cameras capturing the five types of interactions among 13 participants. The data was processed using a DL model with a fine-tuned ResNet (Residual Networks) architecture based on 2D Convolutional Neural Network (CNN) layers for feature extraction. Subsequently, machine learning models were trained and utilized for interaction classification using six commonly used ML algorithms, including SVM, KNN, RF, DT, NB, and XGBoost. The results demonstrate a high accuracy of 95.45% in classifying human interactions. The hybrid approach enabled effective learning, resulting in highly accurate performance across different interaction types. Future work will explore more complex scenarios involving multiple individuals based on the application of this architecture.

## KEYWORDS

Convolutional neural network; deep learning; human interaction recognition; ResNet; skeleton joint key points; human pose estimation; hybrid deep learning and machine learning

## 1 Introduction

In the recent decade, the recognition of human activities is one of the important issues in surveillance video [1–3], human-computer interaction [4,5], video classification [6,7], and several other

fields. The primary objective of human activity recognition (HAR) is to recognize multiple activities automatically from certain videos. To simplify the problem and allow activities to be recognized efficiently, activity recognition was divided into two types: HAR and HIR. Several researchers have focused on HAR, such as rolling, walking, and running, whereas interaction recognition involves the participation of more than one person, for example, pointing, hugging, or pushing. HIR requires more intelligent analysis to distinguish between the multiple interactions that occur among humans. In action recognition, most of the contributions are based on RGB (Red Green Blue) [8], optical flow [9], and skeleton-based data [10,11]. However, action recognition based on RGB and optical flow features requires more extensive computation than skeleton-based features. RGB and optical flow data are used to extract features from the objects of interest, whereas skeleton-based features involve key points rather than the entire object, making them more effective and computationally inexpensive. The prediction of key points in a human skeleton was more reliable and accurate for estimating body poses. Several researchers have worked to improve and develop these body pose estimators [12–14].

The objective of this research was to provide a robust system to recognize human interaction in video sequences captured by multiple cameras, including hugging, kicking, pushing, and pointing. This endeavor was expected to significantly enhance overall human interaction and behavior satisfaction.

### *1.1 Major Contributions*

This study aimed to develop an advanced system for accurately classifying complex human interactions to enhance security, improve human-computer interaction, and create a model that analyzes human-to-human interactions, instead of using the traditional anomaly detection approach, we directly collected prosocial and antisocial behaviors and developed a model to classify these actions. The research makes several key contributions: it utilizes a multi-camera setup to gather a comprehensive dataset that captures human interactions from the front, side, and back, offering a richer perspective than previous studies. Additionally, the study introduces a fine-tuned version of the ResNet model specifically optimized for feature extraction in human interaction scenarios, improving the model's ability to capture subtle and complex features in video data. Furthermore, the research advances the state-of-the-art by focusing on innovative key point extraction from multiple human bodies involved in complex interactions, particularly in scenarios with multiple individuals, where accurate identification and tracking of key points are crucial. By integrating these contributions, the study employs a hybrid approach that combines deep learning (DL) and machine learning (ML) models, including support vector machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), and Extreme Gradient Boosting (XGBoost), to enhance learning effectiveness, accuracy, and generalizability, ensuring that the proposed method outperforms traditional approaches. The application of our study lies in enhancing human behavior analysis across various domains, including surveillance and healthcare by accurately classifying complex human interactions.

The remainder of this paper is organized as follows: Section 2 presents a comprehensive review of the literature on DL and ML human interaction classification. Section 3 elaborates on the methods and data acquisition procedures. Results were detailed in Section 4, followed by a discussion in Section 5.

## 2 Related Work

Human interaction's relevance and impact lie in fostering connection, understanding, and facilitating emotional support, collaboration, and mutual growth. Several studies, such as those by Zhao et al. [15] and Jalal et al. [16], have established specific social dynamics, behavioral analyses, and group activities by focusing on human-human interactions (HHI) in multi-person scenes captured by wearable cameras. However, these studies contain dataset limitations that must be addressed to fully understand the complexities of social dynamics and behaviors. Puchała et al. [17] developed a robust and efficient system to recognize and classify human interactions in video footage using skeleton data and LSTM (Long Short-Term Memory) methods. However, the study's model may miss fine actions, and its reliance on specific datasets could limit broader applicability.

In computer vision, classification tasks are crucial and have diverse applications in fields such as action classification, facial expression classification, and HIR. Yun et al. [18] conducted real-time interaction detection using body-pose features and a dataset that includes synchronized video, depth, and motion capture data. The proposed system used the SVM method for classification. However, the study indicated limitations in scenarios with poor data quality. Ye et al. [19] performed human interaction recognition using manual annotations, proposing methods that utilized a multi-feature fusion network with inception and ResNet. However, this study also had dataset limitations. Shamsipour et al. [20] recognized human interactions in unconstrained videos taken from cameras and remote sensing platforms like drones, which presented a challenging problem. The proposed system used CNN for feature extraction and SVM for classification. The model may struggle in fast-changing environments due to limited frame analysis. Ahad et al. [21] introduced a method to address issues such as motion blur, poor-quality videos, occlusions, variations in body structure or size, and high cognitive or memory demands. The proposed system used the SVM method for interaction classification, but performance was limited, yielding low results with the UTA and SBU datasets.

In recent years, significant advancements have been made in DL-based HIR approaches. Lee et al. [22] combined implicit and explicit representations of human interactions by integrating local image information and primitive motion with body posture. The proposed system used CNN and LSTM methods for classification, but the study faced limitations with the dataset. Stergiou et al. [23] proposed automated recognition of human interactions from video signals, using vision-based methods, particularly DL and CNNs, to tackle these challenges. Wang et al. [24] developed a system for HHI detection using the SaMFormer method, though the performance results were limited. Yin et al. [25] introduced a two-stream hybrid CNN-Transformer network (THCT-Net) to capture local specificity and model global dependencies, respectively. However, this study also encountered dataset constraints. Shafiqul et al. [26] designed a system to detect and recognize human interactions utilizing a Graph Neural Network (GNN) framework.

Moreover, a multi-stream network was proposed for HIR, which combined information from two streams. Haroon et al. [27] suggested a two-stream method, initially using 1D CNN with LSTM and then 3D CNN to recognize human interactions. However, the study highlighted that the model's accuracy could be affected by lighting, background clutter, and variations in clothing or skin tones. Men et al. [28] proposed a two-stream recurrent neural network to recognize HHI from skeletal sequences using the LSTM method to capture temporal properties. Nonetheless, the study had dataset constraints. Furthermore, in the final study, Hachiuma et al. [29] employed LSTM-based models and knowledge-aware feature extraction from skeleton data. Puchała et al. [17] evaluated three models: SC-LSTM-PS (Single Channel Long Short-Term Memory-polar sparse), DC-LSTM-LA (Double

Channel Long Short-Term Memory-Limb-angle features), TC-LSTM-LA (Triple channel Long Short-Term Memory-Limb-angle features), and also faced limitations related to dataset performance.

Existing researches from [15] to [29] primarily focused on single methods like LSTM, SVM, and CNN for human interaction recognition, often constrained by limited datasets and specific scenarios. Studies like Zhao et al. [15] and Jalal et al. [16] emphasized single-person scenes, while others like Puchała et al. [17] and Ye et al. [19] struggled with fine action recognition and data quality. Advanced models, such as those by Lee et al. [22] and Yin et al. [25], integrated DL techniques but were limited by dataset specificity. In contrast, our study introduces a multi-camera setup with a fine-tuned ResNet for feature extraction, coupled with a hybrid approach combining DL and ML methods. This ensures superior accuracy, generalizability, and effectiveness across diverse applications.

## 3 Methods

### *3.1 Dataset Acquisition*

The experiment was conducted on the first floor of the Hak-Ye building (room H111) at Soonchunhyang University. In this study, the HIR experimental approach involved capturing a dataset of human interactions using multiple camera channels. The participants engaged in interactions within a child playroom environment (indoor setting) that was monitored by four cameras. Within the scope of our research, five common classes of human interaction were identified: hugging, kicking, pushing, pointing, and no interaction. This classification is based on the analysis that actions such as hugging for comfort and pointing for assistance represent prosocial behavior, whereas kicking and pushing are indicative of antisocial behavior. The dataset comprised 1.7 s of video recordings for each of the five interaction categories. Four cameras were positioned at different angles in the recording room. The cameras recorded video at a resolution of $1582 \times 1080$ pixels at 30 frames per second, resulting in 51 frames per 1.7 s clip. In this experiment, two participants were selected to enter the room, and then the interactions commenced. To generate a comprehensive dataset for classifying human interactions, each camera (labeled camera 1 through camera 4) captured unique angles and perspectives of the interactions within the room.

The process involved informing the participants about the interactions and then presenting them with images corresponding to each scenario. Subsequently, the participants were prompted to respond to each image by engaging in the corresponding interaction. Recordings were initiated once the interaction began and were stopped when it ceased. This procedure was repeated for multiple images depicting the same interaction, and the interactions were then saved. The process was then repeated for the remaining interaction types. Furthermore, the data collection process provided clarity and insight into the workflow. A detailed flowchart, as depicted in Fig. 1, outlines the sequential steps involved in the process from initiation to conclusion, including informing the participants about the interactions, presenting them with corresponding images, recording their responses, and saving the data.

In this study, 13 participants participated in interactions for a total duration exceeding three hours. There were 10 males and 3 females, aged between 24 and 28 years, with heights ranging from 160 to 180 cm, and weights ranging from 45 to 80 kg. For each interaction, the duration was 5 min to collect data. These interactions were repeated for all participants, with each interaction lasting 20 min. Additionally, each interaction had characteristics such as front hugging, back hugging, front pushing, back pushing, front pointing, back pointing, front kicking, back kicking, and none-interaction (two persons focused on their respective objects), as shown in Fig. 2.

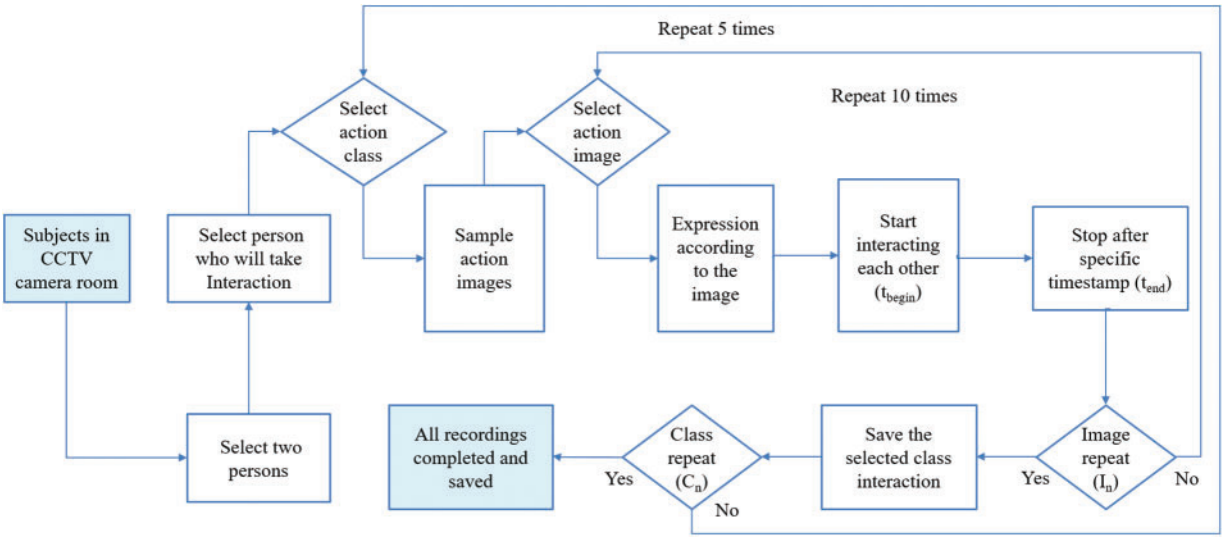


**Figure 1:** Overview of the entire flowchart of data collection

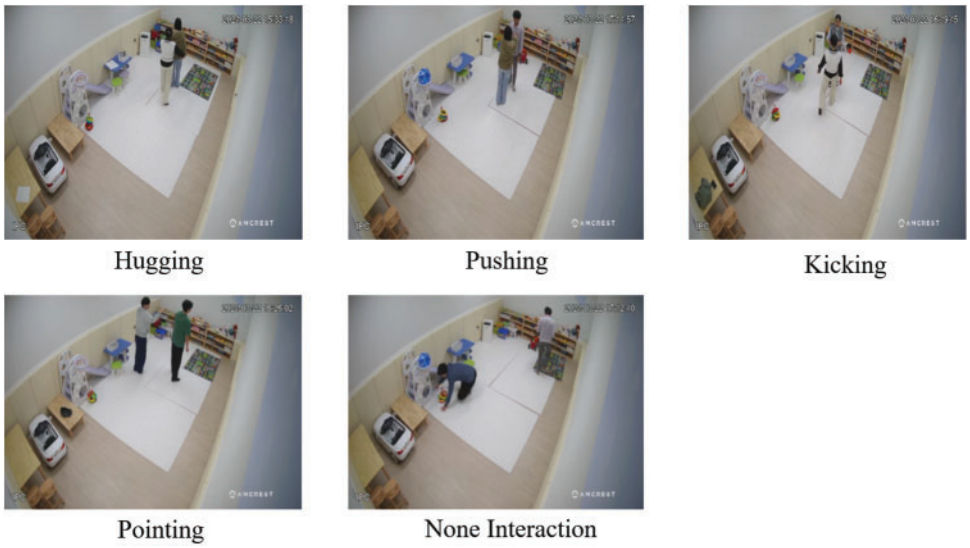


**Figure 2:** Human interaction dataset with five classes

The total number of videos in the dataset after the gathering step for more detailed information is shown in Table 1. Our dataset comprised 550 videos, representing a diverse range of human interactions. The dataset included five classes: hugging, pointing, kicking, pushing, and none interaction. Each class had 110 videos, with each video duration of 1.7 s, 30 frames per second, and a frame resolution of 1582 × 1080 pixels. This dataset was dedicated to human interaction analysis and was instrumental in the development and evaluation of the proposed method, which was specifically recognized for human interaction tasks.

**Table 1:** Representatives of the data collection

| Class | Number of video files | Durations | FPS | Frame resolution | Camera name |
|---|---|---|---|---|---|
| Hugging | 110 | 1.7 s for each interaction | 30 | 1582 × 1080 | 4 cameras |
| Pushing | 110 | | | | |
| Kicking | 110 | | | | |
| Pointing | 110 | | | | |
| None interaction | 110 | | | | |
| **Total** | **550** | | | | |

### *3.2 Data Division*

Data division is a fundamental step in the preparation of datasets for ML and DL tasks. In this study, the dataset, comprised of combining four channel cameras, was split into two main subsets: 80% for training, which was divided into 440 videos, and 20% for testing, which was divided into 110 videos. This division provides a good balance: 80% of the data allows the model to learn effectively, while 20% is enough to test its performance and generalization on unseen data.

### *3.3 Research Methodology*

Hybrid DL and ML classifier methods were proposed to classify human interactions from a sequence of frames based on key features extracted using a pose estimation algorithm. There were four significant blocks: preprocessing, sequence key point extraction, extracted features, and ML classifiers. First, a sequence of frames from the video sequences was extracted. Second, pose estimation extracts human body key points from two-person interactions. Each participant obtained 25 key points. Third, the proposed algorithm employs DL using a ResNet model based on 2D convolutional layers to extract features from human interactions from sequences of frames. Fourth, various ML algorithms were used to classify human interactions based on features extracted from the refined ResNet model. Fig. 3 shows each block of the proposed method.

#### *3.3.1 Preprocessing*

The initial steps involved the consumption of video data, followed by frame extraction, which served as the foundational preprocessing phase. The frames were extracted and resized from the videos. The video was loaded using the OpenCV (Open Source Computer Vision) library, and each frame was read in a loop and resized to 800 × 500.

#### *3.3.2 Sequence Key Point Extraction*

The main purpose of human pose estimation is to localize the key points of the human body (the nose, neck, shoulder, hip, knee, ankle, eye, ear, big toe, small toe, heel, wrist, and elbow) [30] and many applications have been presented in various domains, such as human-computer interaction [31] and the recognition of human activities [32] and pose [33]. Recently, numerous studies have developed based on deep neural networks to achieve the highest performance [34,35]. In the proposed method, interactions, like other actions, constitute sequential data over a period. Instead of using all frames in the videos, 51 frames were extracted from all videos. This approach aimed to reduce the load and

processing time while still capturing the essential motion dynamics in the video data. The frames were downsized to 800 pixels in width and 500 pixels in height to reduce the computational load during the pose estimation process. Pose estimation was used to extract the 2D joint locations of the skeleton. Skeleton key point detection was performed using OpenPose [36]. The proposed OpenPose extracted 25 key points from the human body, where each point represented by 2D X- and Y-coordinate. Fig. 4 shows the results of 50 human skeleton key points generated using a multiple-person pose estimation method based on the OpenPose framework.

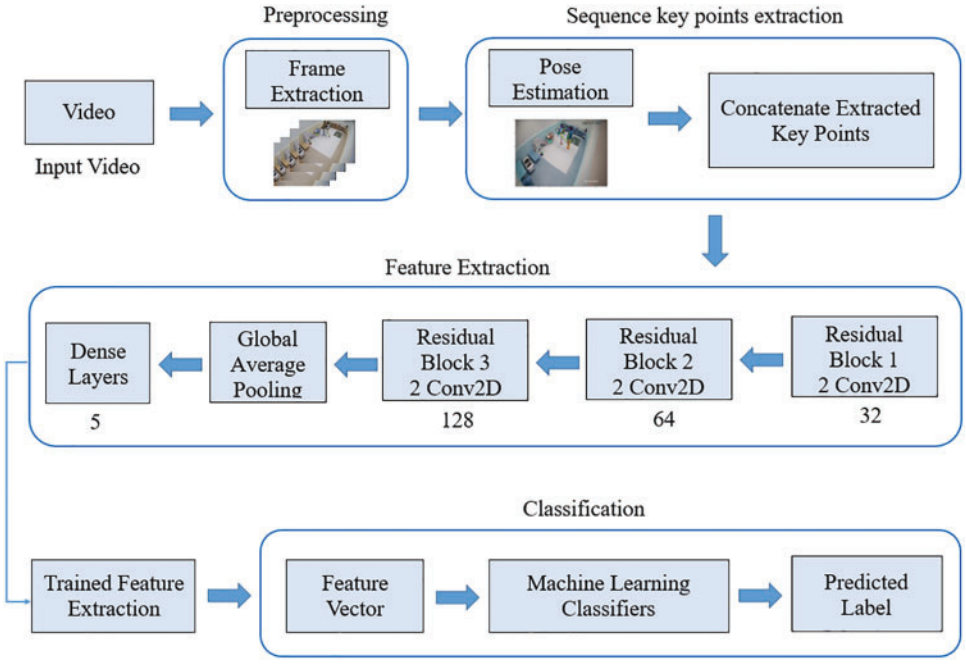


**Figure 3:** Flowchart of the proposed HIR method

(a) Pushing (b) Kicking

**Figure 4:** Samples of skeleton key points in multiple-person pose estimation

The output from OpenPose was an array with a shape of 50 × 3 for detecting two objects in a frame, where 50 represents the number of key points detected for each person and 3 represents the X-coordinate, Y-coordinate, and Z-confidence scores. The Z-confidence score was not used, and it was removed from the array. Therefore, the frames became an array with a shape of 50 × 2, where X values ranged from 0 to 800, corresponding to the width of the frame, and Y values ranged from 0 to

500, corresponding to the height of the frame. The extracted key points for each frame were stacked to construct sequential data. Each sequential datum had a shape of 51 × 50 × 2, where 51 denotes the number of frames, 50 denotes the number of key points detected by each person in each frame, and 2 represents the X- and Y-coordinate of each key point. It was also noted that if fewer than 50 key points were generated for each human, a zero value was assigned to the missing joints. When OpenPose failed to detect any key points for a person, the output was a null value. To address this issue, a zero value was assigned to align with the model input.

### *3.3.3 Feature Extraction*

In this analysis, the ResNet model was fine-tuned to effectively capture hierarchical features and reduce the vanishing gradient problem using residual connections. This model comprised three residual blocks with progressively increasing filter sizes of 32, 64, and 128. Each block featured two convolutional layers with 3 × 3 filters and a stride of 1, followed by batch normalization and the rectified linear unit activation function. These residual blocks incorporate shortcut connections with 1 × 1 convolutions ensuring the preservation of the original input and enhancing gradient flow. Between the residual blocks, max-pooling layers were used to down sample the spatial dimensions, while dropout layers supported regularization and prevented overfitting. The comprehensive architecture integrates 32 million trainable parameters. Within the scope of this study, modifications were made to the ResNet model. Specifically, the requirements of a human interaction dataset comprising five distinct classes were addressed. The fine-tuned model was then suitable for extracting features from the global average pooling 2D layer yielding feature vectors characterized by dimensions. A visual representation of this adapted model's architecture is provided in Fig. 5.

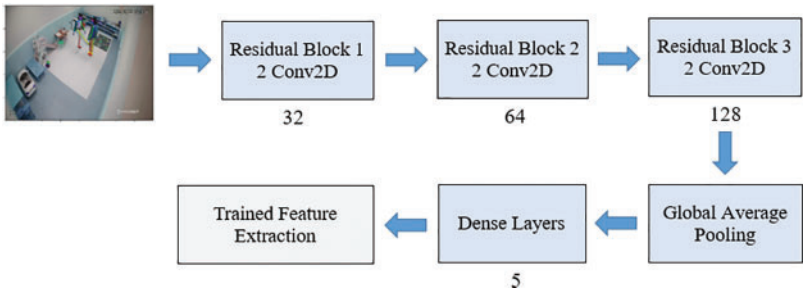


**Figure 5:** Model architecture of feature extraction on HIR

In addition, this vector was passed through two dense layers. The first dense layer consisted of 100 with the rectified linear unit activation function, and the final dense layer had five output vectors. The output comprises five vectors representing the distribution class probabilities obtained using the SoftMax activation function. The trained feature extraction layer is then used to obtain the final feature vectors for further interaction classification.

### *3.3.4 Classification*

A DL model using the fine-tuned ResNet architecture was implemented to extract significant features from video sequences. The features were extracted at the global average pooling 2D layers. The extracted features of the proposed method were converted into a feature vector and then passed to six popular ML classifiers, including SVM, KNN, RF, DT, NB, and XGBoost [37]. Various ML classifiers used different techniques for learning, determined five distinct classes: hugging, kicking, pushing, pointing, and none interaction of our dataset, and improved the performance results.

We selected the top two highest-performing ML models (SVM and KNN) from our research that include a detailed description of mathematical equations. For each method, provide a concise explanation of the key mathematical concepts and equations. First, SVM includes the equation for the decision boundary (Eq. (1)) where ω is weight, $x$ feature vector, and b is the bias as in the following formula:

$$\omega \cdot x + b = 0 \tag{1}$$

Second, KNN include the equation of distance calculation (Eq. (2)) where $x$ and $x'$ are feature vectors of two data points, and $n$ is the number of features.

$$d(x,\ x') = \sqrt{\sum_{i-1}^{n} (x_i - x_i')^2} \tag{2}$$

### *3.4 Model Training*

In our experiments, we selected the sparse categorical cross entropy for this architecture, which was appropriate for multiple-class classification tasks where the labels were integers while still providing an effective measure of loss for training DL models.

In the optimizer, Adam was selected for this architecture, primarily, due to its reputation for facilitating rapid convergence and its robustness in handling complex optimization problems. A batch size of 4 was chosen, which was suitable for our system. The learning rate was another crucial hyperparameter that influenced how the neural network learned. A rate of 0.001 was selected to balance effective learning and stability. For our classification model, we fine-tuned a 2D CNN of the ResNet architecture using Tensor Flow, Keras, OpenCV, Numpy, and Scikit-Learn in Python programming language, version 3.10.3. The training was conducted on a desktop with an NVIDIA TITAN RTX GPU (Graphic Processing Unit), Intel Xeon Silver 4114 CPUs (Central Processing Unit), and 192 GB of RAM (Random Access Memory).

The training process of the proposed DL model was conducted over 300 epochs. An epoch is defined as a single complete pass through the entire training dataset. The number of epochs was a vital hyperparameter in DL because it determined how many times the model updated its weights based on the training data. The refined configuration of the fine-tuned hyperparameter is summarized in Table 2.

**Table 2:** Summary of fine-tuning parameters during training

| Hyperparameter | Value |
|---|---|
| DL framework | Tensor flow |
| Optimizer | Adam |
| Activation function | ReLU (Rectified Linear Unit) |
| Learning rate | 0.001 |
| Dropout | 0.2 |
| Batch size | 4 |
| Epoch | 300 |
| Frame size | 800 × 500 |

## 4 Results

### *4.1 Experimental Results*

An experiment was conducted using the proposed hybrid DL and ML classifier models on our dataset. Multiple classifiers were assessed across a range of performance metrics, including precision rate, recall rate, F1 score, and accuracy. The best performances of the four classifiers, including their classification results and the confusion matrix, are presented in this section. The confusion matrix was examined to gain insights into the model's classification outcomes.

The results of applying the proposed method using DL and six ML classifiers to classify the human interaction dataset are presented in Table 3. The implementation of the proposed method aimed to validate its performance in classifying human interactions, achieving an overall accuracy of 95.45%. First, SVM achieved performance metrics with a recall rate of 0.96, a precision rate of 0.96, and an F1 score of 0.96. Second, KNN showed a recall rate of 0.95, a precision rate of 0.95, and an F1 score of 0.95. Third, DT obtained a recall rate of 0.94, a precision rate of 0.94, and an F1 score of 0.94. Fourth, NB recorded a recall rate of 0.92, a precision rate of 0.93, and an F1 score of 0.92. The computational time for this evaluation was 1396.83 s. Furthermore, the confusion matrix for the top four results using our dataset is shown in Fig. 6. Fig. 7 shows the receiver operating characteristic curves of HIR.

**Table 3:** Results of the proposed method using hybrid DL with six ML classifiers

| Approach | ML classifier | Precision | Recall | F1 score | Accuracy |
|---|---|---|---|---|---|
| Fine-Tuned ResNet | SVM | 0.96 | 0.96 | 0.96 | **95.45** |
| | KNN | 0.95 | 0.95 | 0.95 | **95.45** |
| | DT | 0.94 | 0.94 | 0.94 | **93.63** |
| | NB | 0.93 | 0.92 | 0.92 | **91.81** |
| | RF | 0.91 | 0.91 | 0.91 | 90.90 |
| | XGBoost | 0.84 | 0.82 | 0.81 | 81.81 |

### *4.2 Evaluation*

The evaluation of the proposed model was an essential part of this study, where we assessed its performance on the target dataset. The primary objective of this evaluation was to determine the model's effectiveness in terms of accurately classifying and recognizing relevant features in this dataset. To realize this, we used a range of performance metrics tailored to the recognition task. Specifically, we focused on accuracy, precision, recall, and the F1 score as key evaluation criteria.

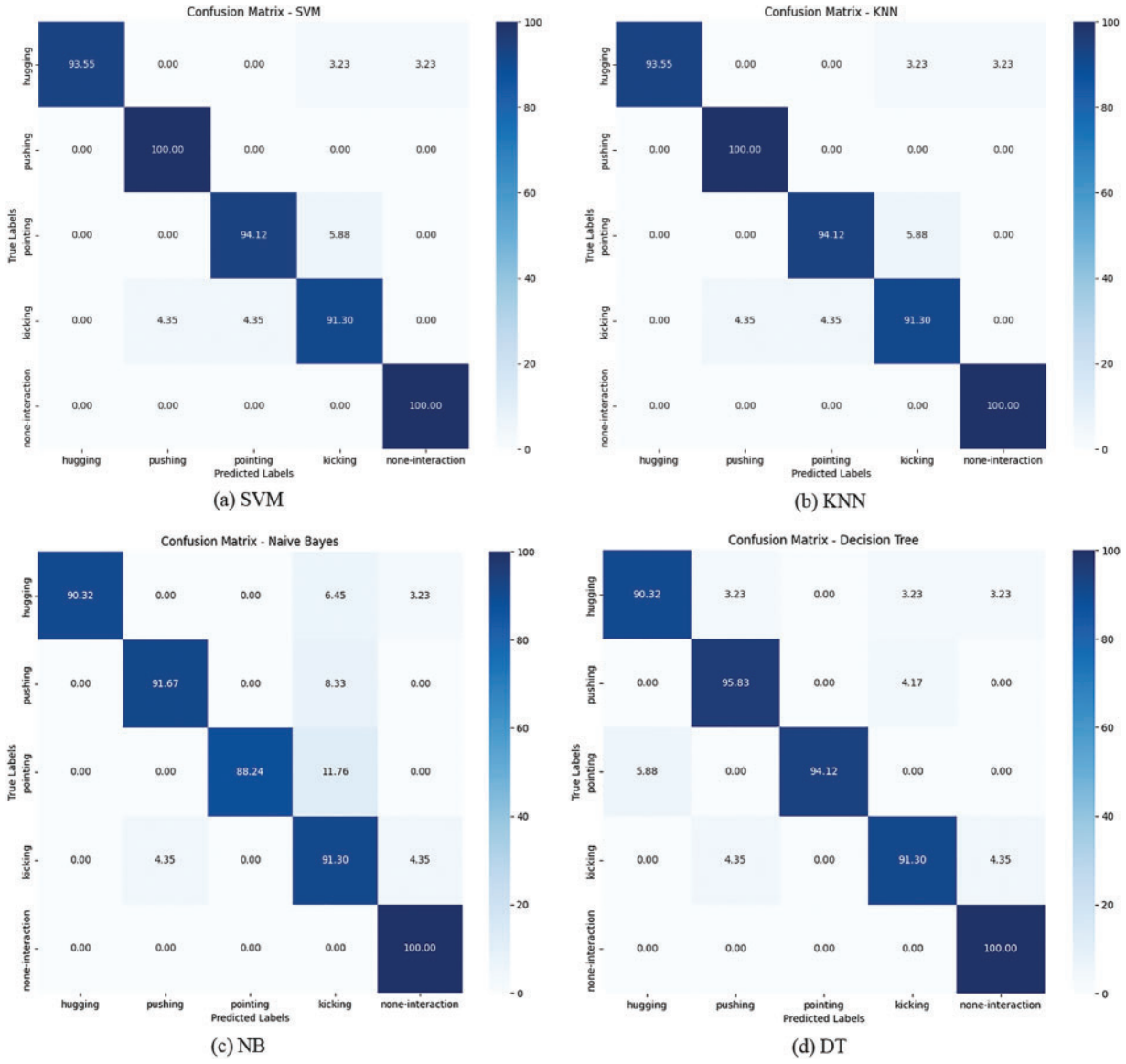


**Figure 6:** Confusion matrix of the proposed hybrid method using DL with SVM, KNN classifiers, NB and DT classifiers

As shown in Table 4, our proposed method based on hybrid DL and ML showed an overall accuracy of 95.45%, a recall rate of 0.96, a precision rate of 0.96, and an F1 score of 0.96. The computational time of this evaluation was 1396.83 s. Performance results from another experiment are also presented in this table.

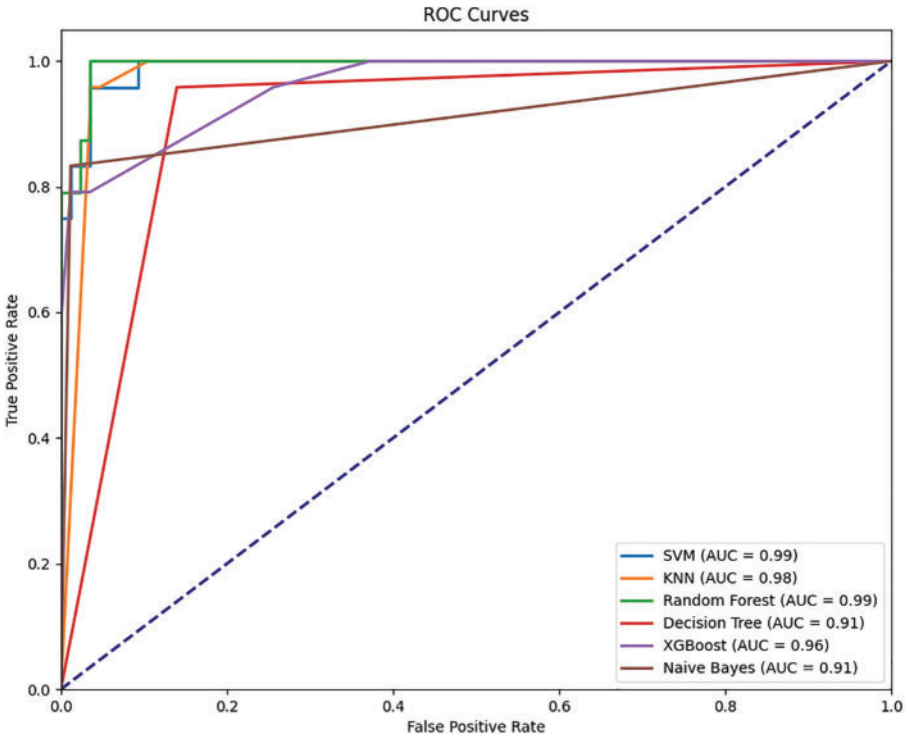


**Figure 7:** Receiver operating characteristic curves of HIR

**Table 4:** Comparison of the proposed method with related methods on HIR in terms of classification accuracy

| Approaches | Accuracy |
|---|---|
| **Our proposed** | **95.45** |
| SVM [19] | 91 |
| GNN [15] | 89.3 |
| CNN + SVM [20] | 90.42 |
| SVM [21] | 74.40 |
| CNN-Transformer [25] | 91.00 |

## 5 Conclusion

This paper has presented a hybrid approach for intelligently classifying various human interactions. To analyze the patterns generated from human poses, frames were extracted, and the human body was localized over a sequence of frames to determine the positions of the skeleton's joints. This skeletal information was then input into a fine-tuned ResNet model, which efficiently extracted discriminative features and learned complex sequences of patterns. These features were transferred to a feature vector and passed through popular ML classifiers, accurately discriminating between the various human interactions observed in the videos. Additionally, experiments were conducted

using input key points directly with the ML model, bypassing the feature extraction step. Results demonstrated lower performance in this scenario compared to using the feature extraction approach. Experiments verified that the proposed network achieved excellent results on the human interaction dataset.

In the future, the proposed hybrid DL and ML classifier architecture will be further explored by focusing on individual humans. Additionally, we plan to incorporate feature encoding to enhance the model's ability to capture and represent human interactions more effectively. This approach will allow the robust interrelated dynamics of each human to be inferred using the proposed hybrid network, leading to even higher performance in recognition tasks involving more complex and multiple human interactions.

**Acknowledgement:** The authors would like to express our sincere gratitude and appreciation to each other for our combined efforts and contributions throughout the course of this research paper.

**Funding Statement:** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176) and the Soonchunhyang University Research Fund.

**Author Contributions:** Study conception and design: Vesal Khean, Chomyong Kim, Sunjoo Ryu; analysis and interpretation of results: Awais Khan, Min Kyung Hong, Eun Young Kim, and Joungmin Kim; draft manuscript preparation: Yunyoung Nam. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** Not applicable.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] W. Ullah, A. Ullah, T. Hussain, Z. A. Khan, and S. W. Baik, "An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos," *Sensors*, vol. 21, no. 8, 2021, Art. no. 2811. doi: 10.3390/s21082811.

[2] A. Ullah, K. Muhammad, W. Ding, V. Palade, I. U. Haq and S. W. Baik, "Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications," *Appl. Soft Comput.*, vol. 103, 2021, Art. no. 107102. doi: 10.1016/j.asoc.2021.107102.

[3] W. Ullah *et al.*, "Artificial Intelligence of things-assisted two-stream neural network for anomaly detection in surveillance big video data," *Future Gener. Comput. Syst.*, vol. 129, pp. 286–297, 2022. doi: 10.1016/j.future.2021.10.033.

[4] S. Ahmed, K. D. Kallu, S. Ahmed, and S. H. Cho, "Hand gestures recognition using radar sensors for human-computer-interaction: A review," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 527. doi: 10.3390/rs13030527.

[5] L. Xu, J. Hou, and J. Gao, "A novel smart depression recognition method using human-computer interaction system," *Wirel. Commun. Mob. Comput.*, vol. 2021, no. 1, 2021, Art. no. 5565967. doi: 10.1155/2021/5565967.

[6] F. U. M. Ullah, M. S. Obaidat, K. Muhammad, and A. Ullah, "An intelligent system for complex violence pattern analysis and detection," *Int. J. Intell. Syst.*, vol. 37, no. 12, pp. 10400–10422, 2022. doi: 10.1002/int.22537.

[7] L. Wang, H. Zhang, and G. Yuan, "Big data and deep learning-based video classification model for sports," *Wirel. Commun. Mob. Comput.*, vol. 2021, no. 1, 2021, Art. no. 1140611. doi: 10.1155/2021/1140611.

[8] N. Soans, E. Asali, Y. Hong, and P. Doshi, "Sa-Net: Robust state-action recognition for learning from observations," in *2020 IEEE Int. Conf. Robot. Autom. (ICRA)*, 2020, pp. 2153–2159.

[9] H. Kim, S. Park, H. Park, and J. Paik, "Enhanced action recognition using multiple stream deep learning with optical flow and weighted sum," *Sensors*, vol. 20, no. 14, 2020, Art. no. 3894. doi: 10.3390/s20143894.

[10] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D pose estimation and action recognition using multitask deep learning," in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 5137–5146.

[11] D. Ludl, T. Gulde, and C. Curio, "Enhancing data-driven algorithms for human pose estimation and action recognition through simulation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3990–3999, 2020. doi: 10.1109/TITS.2020.2988504.

[12] M. Segu, F. Pirovano, G. Fumagalli, and A. Fabris, "Depth-aware action recognition: Pose-motion encoding through temporal heatmaps," 2020, *arXiv:2011.13399*.

[13] Z. Cao, T. Simon, S. -E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 1302–1310.

[14] C. Wang, F. Zhang, and S. S. Ge, "A comprehensive survey on 2D multi-person pose estimation methods," *Eng. Appl. Artif. Intell.*, vol. 102, 2021, Art. no. 104260. doi: 10.1016/j.engappai.2021.104260.

[15] J. Zhao, R. Han, Y. Gan, L. Wan, W. Feng and S. Wang, "Human identification and interaction detection in cross-view multi-person videos with wearable cameras," in *Proc. 28th ACM Int. Conf. Multimed.*, Seattle, WA, USA, Oct. 12–16, 2020, pp. 2608–2616.

[16] A. Jalal, N. Khalid, and K. Kim, "Automatic recognition of human interaction via hybrid descriptors and maximum entropy markov model using depth sensors," *Entropy*, vol. 22, no. 8, 2020, Art. no. 817. doi: 10.3390/e22080817.

[17] S. Puchała, W. Kasprzak, and P. Piwowarski, "Human interaction classification in sliding video windows using skeleton data tracking and feature extraction," *Sensors*, vol. 23, no. 14, 2023, Art. no. 6279. doi: 10.3390/s23146279.

[18] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *2012 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Providence, RI, USA, 2012, pp. 28–35.

[19] Q. Ye, H. Zhong, C. Qu, and Y. Zhang, "Human interaction recognition based on whole-individual detection," *Sensors*, vol. 20, no. 8, 2020, Art. no. 2346. doi: 10.3390/s20082346.

[20] G. Shamsipour and S. Pirasteh, "Artificial intelligence and convolutional neural network for recognition of human interaction by video from drone," 2019, *arXiv:201908.0289.v1*.

[21] M. A. R. Ahad and T. Paul, "Human-human interaction recognition based on gradient-based features," *Int. J. Biomed. Soft Comput. Human Sci.: Official J. Biomed. Fuzzy Syst. Assoc.*, vol. 25, no. 2, pp. 39–47, 2020.

[22] D. -G. Lee and S. -W. Lee, "Human interaction recognition framework based on interacting body part attention," *Pattern Recognit.*, vol. 128, 2022, Art. no. 108645. doi: 10.1016/j.patcog.2022.108645.

[23] A. Stergiou and R. Poppe, "Analyzing human–human interactions: A survey," *Comput. Vis. Image Understanding*, vol. 188, 2019, Art. no. 102799. doi: 10.1016/j.cviu.2019.102799.

[24] Z. Wang, K. Ying, J. Meng, and J. Ning, "Human-to-human interaction detection," in *Int. Conf. Neural Inform. Process.*, Changsha, China, Nov. 20–23, 2023, pp. 120–132.

[25] R. Yin and J. Yin, "A two-stream hybrid CNN-transformer network for skeleton-based human interaction recognition," 2023, *arXiv:2401.00409*.

[26] I. M. Shafiqul, M. K. A. Jannat, J. -W. Kim, S. -W. Lee, and S. -H. Yang, "HHI-AttentionNet: An enhanced human-human interaction recognition method based on a lightweight deep learning model with attention network from CSI," *Sensors*, vol. 22, no. 16, 2022, Art. no. 6018. doi: 10.3390/s22166018.

[27] U. Haroon *et al.*, "A multi-stream sequence learning framework for human interaction recognition," *IEEE Trans. Hum.-Mach. Syst.*, vol. 52, no. 3, pp. 435–444, 2022. doi: 10.1109/THMS.2021.3138708.

[28] Q. Men, E. S. Ho, H. P. Shum, and H. Leung, "A two-stream recurrent network for skeleton-based human interaction recognition," in *2020 25th Int. Conf. Pattern Recognit. (ICPR)*, Milan, Italy, 2021, pp. 2771–2778.

[29] R. Hachiuma, F. Sato, and T. Sekii, "Unified keypoint-based action recognition framework via structured keypoint pooling," in *2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 22962–22971. doi: 10.1109/CVPR52729.2023.02199.

[30] E. Ng, D. Xiang, H. Joo, and K. Grauman, "You2Me: Inferring body pose in egocentric video via first and second person interactions," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 9887–9897.

[31] S. Wu, H. Kan, J. Gao, and W. Yue, "Convolutional neural networks-motivated high-performance multi-functional electronic skin for intelligent human-computer interaction," *Nano Energy*, vol. 122, 2024, Art. no. 109313. doi: 10.1016/j.nanoen.2024.109313.

[32] M. A. Khan *et al.*, "Human action recognition using fusion of multiview and deep features: An application to video surveillance," *Multimed. Tools Appl.*, vol. 83, no. 5, pp. 14885–14911, 2024. doi: 10.1007/s11042-020-08806-9.

[33] H. Chen, R. Feng, S. Wu, H. Xu, F. Zhou and Z. Liu, "2D human pose estimation: A survey," *Multimed. Syst.*, vol. 29, no. 5, pp. 3115–3138, 2023. doi: 10.1007/s00530-022-01019-0.

[34] B. Artacho and A. Savakis, "UniPose: Unified human pose estimation in single images and videos," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 7033–7042.

[35] R. Tahir and Y. Cai, "Multi-human pose estimation by deep learning-based sequential approach for human keypoint position and human body detection," *J. Shanghai Jiaotong Univ. (Sci.)*, vol. 28, no. 3, pp. 1–11, 2023. doi: 10.1007/s12204-023-2658-z.

[36] S. -E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 4724–4732.

[37] M. C. Untoro, M. Praseptiawan, M. Widianingsih, I. F. Ashari, and A. Afriansyah, "Evaluation of decision tree, k-NN, Naive Bayes and SVM with MWMOTE on UCI dataset," *J. Phys.: Conf. Series*, vol. 1477, no. 3, 2020, Art. no. 032005. doi: 10.1088/1742-6596/1477/3/032005.