A blended ensemble approach for accurate human activity recognition

Rezwana Karim[1], Afsana Begum[1], Miskatul Jannat[2], Abu Kowshir Bitto[1]
[1]Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering, International Islamic University Chittagong, Chittagong, Bangladesh

Article Info

Article history:

Received May 28, 2025
Revised Oct 10, 2025
Accepted Nov 8, 2025

Keywords:

Ensemble model
Human activity recognition
Recognition applications
Scalable vision models
Transformer architectures

ABSTRACT

Human activity recognition (HAR) is a novel computer vision area with applications in fashion, entertainment, healthcare, and urban planning. Previously, convolutional neural networks (CNNs) were used in HAR due to their ability to extract spatial features from images. However, CNNs are not effective in processing varying input sizes and long-range dependencies in complex human motions. This work examines another approach using vision transformers (ViT) and swin transformers (SwinT) that process images as patch sequences and perform self-attention. These models particularly excel in learning global relationships and minor motion changes in body motion and are therefore very well-suited to variegated and subtle activity detection. To further enhance recognition performance, we propose a hybrid ensemble method by combining ViT and SwinT models with different scales (small, base, and large). Experimental outcomes show that while single transformer models are competitive, the hybrid ensemble beats them across the board with the highest accuracy and balanced precision, recall, and F1-score. These findings confirm that the intended ensemble model provides a more scalable and robust solution than either single-model or CNN-based approaches, and this encourages accurate human activity recognition.

Corresponding Author:

Abu Kowshir Bitto
Department of Software Engineering, Daffodil International University
Dhaka-1216, Bangladesh
Email: abu.kowshir777@gmail.com

1. INTRODUCTION

In recent, the potential of machines to automatically detect and classify human activity from visual information has been a topic of immense interest in research and practical applications [1]. Human activity recognition (HAR) is a key technology for numerous applications such as healthcare monitoring, surveillance systems, human-computer interaction, sports analytics, and smart environments. The advancement of video surveillance systems, wearable sensors, and intelligent cameras has brought about an explosive increase in the amount of activity-related data, thereby creating new challenges in the correct interpretation and classification of intricate human activities [2].

HAR systems relied on handmade feature engineering and conventional machine learning methods [3]. Yet, such methods tend to fail in adequately capturing the complex spatial and temporal processes of human activities, especially in heterogeneous and unstructured settings. The advancement of deep learning techniques, specifically convolutional neural networks (CNNs), was a breakthrough in the sense that it enabled automatic feature extraction and representation learning directly from raw visual data [4]. Even though they are successful, CNN-based models are at times inadequate in capturing long-range dependencies and intricate spatial relationships, which are essential for comprehending sophisticated human activities. Most recently, ensemble-based methods CNN-based transfer learning ensembles and early-exit networks, proved that the union of heterogeneous models greatly increased robustness as well as recognition accuracy [5].

In response to these constraints, the computer vision has progressively shifted towards transformer-based models, which were initially popular in natural language processing [6]. Vision transformers (ViT) and their variations have proven exceptionally effective in numerous image recognition applications by efficiently learning global dependencies with self-attention mechanisms. Among them, the swin transformer (SwinT) a hierarchical ViT model has shown a promising architecture by combining the merits of both transformers and CNNs, offering improved efficiency and scalability for dense vision tasks [7].

There have been several studies on sensor-based HAR using accelerometer, gyroscope, and radar sensor data. Huan *et al.* [8] presented a light-weight hybrid ViT network for radar-based HAR, which combined convolutional operations and self-attention for processing micro-doppler maps efficiently. Ullah and Munir [9] present cascade dual attention CNN with a bi-directional gated recurrent unit (GRU) has also been proposed to learn both spatial and temporal features, leading to improved recognition accuracy in the scenario of HAR tasks. These studies indicate the potential for combining transformer models with traditional deep learning approaches to improve the performance of sensor-based HAR systems.

Vaghela *et al.* [10] used feature fusion approaches to enhance activity recognition from multimodal data. Morshed *et. al.* [11] improved recognition by data fusion with feature engineering, showcasing how the incorporation of carefully chosen handcrafted features with machine learning could be used to increase accuracy. Comprehensive surveys on the evolution from handcrafted to deep networks, with opportunities and challenges for the combination of handcrafted and learned features [12]. Ulhaq *et al.* [13] did survey and shows how action recognition advancements through the integration of handcrafted and learned features.

The recent years have seen rapid development towards transformer-based models for HAR. The initial ViT demonstrated impressive image recognition ability at scale by representing global context information showed by Dosovitskiy *et al.* [14]. SwinT improved on this by introducing hierarchical attention mechanisms with shifted windows to improve computational efficiency showed by Liu *et al.* [15]. Wensel *et al.* [16] and Reda *et al.* [17] present architectures such as ViT-recurrent transformer (ReT), having integrated recurrent and ViT modules for more accurate video activity recognition and ConViViT, which combined convolutional layers with factorized self-attention continued to advance spatiotemporal modeling. Han *et al.* [18] present a novel approach using ViT for human activity recognition was presented, taking advantage of the model's strength in capturing large-scale contextual information, hence achieving higher recognition performance. Additionally Wang *et al.* [19] gait recognition has been improved with the introduction of global-local feature fusion using SwinT and 3D CNN, which successfully extracts spatial and temporal features from gait sequences. The research in [20], [21] present hybrid ViT for efficient HAR and a ViT model for action recognition from still images were significant improvement.

The research in [22], [23] investigates the capability of SwinT and ViT models in human activity recognition through their incorporation into a hybrid ensemble learning model. The application of the ensemble approach is to leverage the complementary strengths of different models so as to improve the robustness and generalizability of HAR systems. Through this process, this paper adds to the body of work in activity recognition by assessing transformer models and showing how ensemble methods can be used to improve performance in real-world and challenging conditions.

2. METHOD

This section is organized into five key sub-sections: methodology, dataset, data preprocessing, data visualization, and model description. Each part provides a detailed explanation to give the reader a clear understanding of the overall approach taken in this study. From Figure 1 we can see that this study proposes a robust methodology for HAR using deep learning and efficient data processing. Image data representing various activities is collected and preprocessed by resizing, normalizing pixel values, and one-hot encoding class labels. The dataset is split into training (80%), validation (10%), and testing (10%) sets with balanced class distribution. Transfer learning is applied using pre-trained ViT and SwinT models in small, base, and large variants. These models are fine-tuned on the activity dataset to capture essential spatial-temporal patterns. To leverage the strengths of the models, an ensemble fusion strategy is adopted: the probability estimates from individual models are blended using a stacked learning paradigm where traditional machine learning classifiers (such as support vector classifier (SVC), logistic regression (LR), random forest (RF), gradient boosting (GB), k-nearest neighbor (KNN), and XGBoost (XGB) are employed as meta-learners. The technique ensures that both the deep feature representations and various decision-making approaches are utilized to generate the ultimate prediction. For validating effectiveness, ablation studies compare single-model and ensemble performance ratings, while multiple experimental run statistical tests validate the stability and robustness of the proposed method. Evaluation metrics include accuracy, precision, recall, and F1-score.

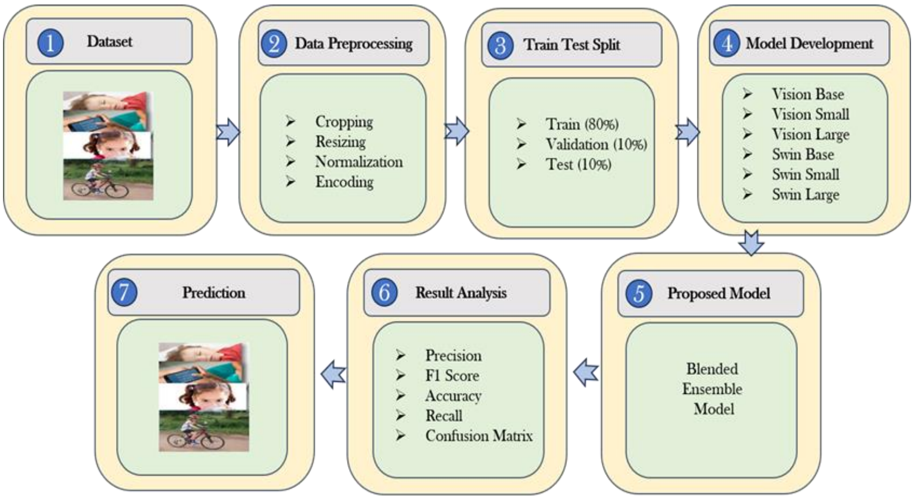

Figure 1. Step by step procudure diagram

2.1. Dataset

To carry out the experiments that are presented in this study, a publicly accessible human action detection dataset obtained from Kaggle [24] was utilized. The dataset has 18,000 labeled images, which are well arranged into 15 different human action classes. The actions within the dataset are calling, clapping, cycling, dancing, drinking, eating, fighting, hugging, laughing, listening to music, running, sitting, sleeping, texting, and using laptop. Every class is filled with exactly 1,000 training images and 200 test images to give a balanced class distribution for robust model testing.15,000 images were assigned for training and the remaining 3,000 images for testing. Further, while training the model, 10% of the respective class training images were also split as the validation set. The validation dataset was utilized to decide on the model performance and tune hyperparameters to prevent overfitting and enhance the generalizability of the proposed method.

2.2. Data preprocessing

Our dataset was well-balanced and free of missing values or noise. They were, however, of varying sizes, which posed an issue for batch processing. We employed a sequence of preprocessing methods to standardize and make them deep learning-friendly. First, we resized all images to 224×224 pixels for consistency in the dataset. Resizing was necessary to facilitate mini-batch training, where images must be the same size. Next, we performed pixel normalization, normalizing pixel values to the [0, 1] range by dividing by 255. This normalization step is important to stabilize and speed up the training process, especially in transfer learning issues. We also one-hot encoded the class labels. The class labels (e.g., "calling," "clapping," "cycling") were initially converted to numerical form (e.g., 0, 1, 2), then converted to binary vectors. For example, for 15 classes, the label "calling" is [1, 0, 0, ..., 0], and so on. These preprocessing methods ensure our data was clean, normalized, and ready for effective model training.

2.3. Proposed ensemble model

Our proposed ensemble model aggregates several pre-trained transformer models to improve the accuracy and stability of the classification [25]. The pipeline begins with a standard data preprocessing phase where all input images are reshaped and resized to 224×224 pixels for uniformity and pixels normalized to scale the values in the range [0, 1]. Class labels are also encoded using one-hot encoding to prepare them for classification. Following preprocessing, the images are passed through six pre-trained models of two kinds: ViT (ViT-B/16, ViT-L/16, and ViT-S/16) and SwinT (swin-B, swin-L, and swin-S). Each model produces feature vectors of varying lengths, which are passed through a dense layer and an activation layer to produce per-category predictions. These then became combined using a stacking ensemble method by feeding in all probability outputs of transformers as input features to an ensemble of different machine learning classifiers. We used SVC, LR, RF, GB, KNN, and XGB as meta-learners. All of these classifiers impart various inductive biases, thereby enabling the meta-layer to learn linear as well as non-linear decision boundaries and enhance generalization. The final predicted class is obtained by consolidating the outputs of the meta-learners, thus gaining from the complementary strengths of both deep transformer models and traditional ensemble classifiers. This hybrid stacking ensemble successfully enhances the performance of the overall classification system. Figure 2 shows our proposed model workflow.

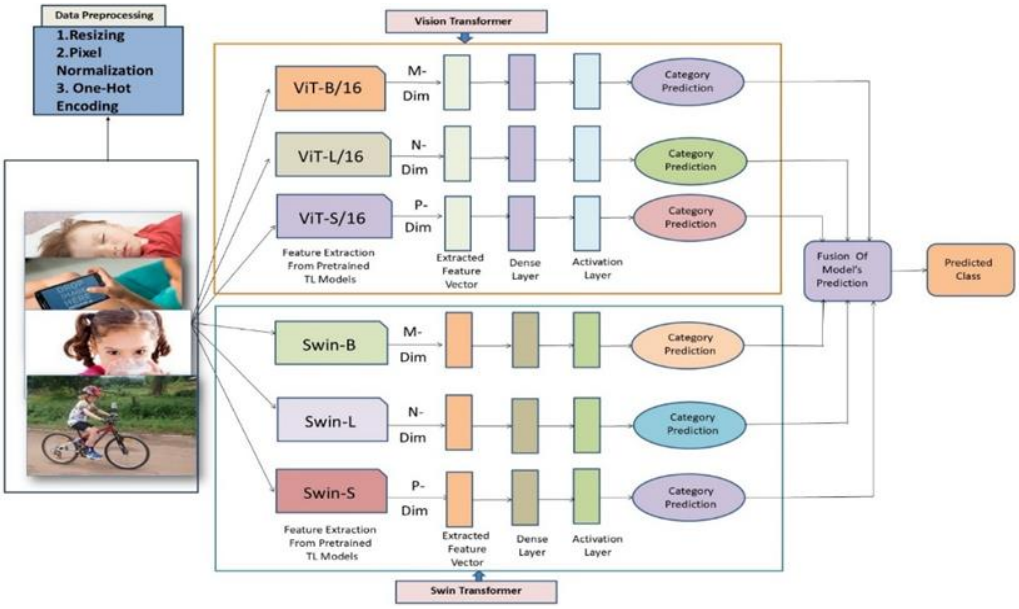

Figure 2. Our proposed ensemble model workflow diagram

2.4. Model performance calculation

To assessing the performance of the proposed HAR model, a comprehensive set of evaluation metrics is employed. As the problem in question is multi-class in nature, evaluation metrics such as accuracy, precision, recall, and F1-score are employed for quantifying the classification accuracy of the model on each of the activity types. These metrics not only allow the quantification of overall correctness but the proportion of correctly predicted activities versus incorrect predictions as well. Moreover, a confusion matrix can be employed to display the model predictions on a per-class basis, providing rich information about what activities are correctly recognized and where misclassifications occur. The following are some of the performance metrics that were calculated. From these parameters we identified the best classifier to recognize HAR. Most of the performance metrics in percentage (%) have been calculated based on (1)-(4) based on the confusion matrix explained in obtained from the classifier.

$$Accuracy = (\frac{TP+TN}{TP+FN+FP+TN}) \times 100\% \tag{1}$$

$$Recall = (\frac{TP}{TP+FN}) \times 100\% \tag{2}$$

$$Precision = (\frac{TP}{TP+FP}) \times 100\% \tag{3}$$

$$F1-score = (2 \times \frac{Precision \times Recall}{Precision+Recall}) \times 100\% \tag{4}$$

3. RESULTS AND DISCUSSION

From Table 1 we can see that ViT base model performs well in terms of some good classification abilities for the 15 activity classes with overall accuracy ranging from 95-98% depending on the activity. It does exceedingly well when identifying individual activities like cycling with near perfect accuracy (96.57%) and recall (98.5%), and running with similarly high rates. However, the model suffers in actions like phone call and clap, where precision (70.52 and 74.73%) and recall (61 and 69.5%) are notably lower, implying misclassifications due to perhaps similar visual features or subtle action dissimilarities. Furthermore, music listening comes with moderate accuracy of 58.37% even with enhanced recall of 71.5%, showing that the model has difficulty separating this class from others as it may have overlapping features with texting or using a laptop. Sitting performance is also comparatively low with accuracy at 51.62%, showing confusion with other still or low-motion activities. This model, although efficient, shows the failing of an unadulterated ViT architecture in totally encompassing subtle activity patterns.

From Table 2 we can see that scaling to ViT large yields performance increases on most activities, with accuracy and recall especially improving for tricky classes like calling (accuracy 76.74%, recall 66%), clapping (accuracy 77.96%, recall 72.5%), and music (accuracy 63.9%, recall 77%). The increased depth and model capacity allow increased discriminative feature learning, which is evident in nearly perfect cycling performance (99.5% precision and 99% recall) and better interaction with dynamic activities such as running and drinking. Nevertheless, certain classes such as fighting see a decrease in precision to 74.38% even with good recall (90%), which may be because of episodic false positives possibly due to overlapping action features. Moderate accuracy on sitting (70.86%) and laptop use (74.25%) also suggests some residual difficulty in distinguishing sedentary behavior. Generally, vision large obtains an equilibrium between greater accuracy and greater class separation, yet still struggles with visually similar or subtle actions.

From Table 3 we can see that the ViT small model, with its smaller parameter size, generally shows lower performance all around, especially on more ambiguous activities. Call and clap precision and recall drop under 70%, with call precision coming in at 61.82%, which suggests decreased capability to differentiate subtle motions. Despite this, the model remains highly functional on clear activities like cycling (98.48% precision) and eating (94.86% precision), demonstrating that easier classes remain well-identified. Interestingly, tasks with dynamic interaction such as fighting suffer in precision (58.96%) while recall remains good (90.5%), implying false positives most likely due to limited model capacity. The reduction in precision and recall for laughter and music also points to difficulty with subtle emotional or background activity. This model may be better deployed in contexts where resources are limited but can afford to make errors on subtle classes.

From Table 4 we can see that the SwinT base model constantly improves classification scores by employing hierarchical representation and local-global attention. Precision and recall are greatly enhanced for most classes; for example, drinking has 92.82% precision and 84% recall, whereas hugging has 90.86% precision and 84.5% recall. Spatial-temporal fine-grained subtleties are effectively captured by the architecture as evidenced by high F1-scores in eating (90.4%) and running (86.46%). There are still difficulties in laughing and texting accurately at around 70% that indicate lingering class confusion, but overall balance of classes is better than in ViT alone. The model also has good recall in dynamic actions such as combat (85.5%) and cycling (99%), so it can be a good option for applications where subtle recognition is needed with tolerable computational load.

Table 1. Performance of ViT base models

Class	TP	FP	FN	TN	Precision	Recall	F1-score	Accuracy
calling	122	51	78	2749	70.52	61.00	65.42	95.70
clapping	139	47	61	2753	74.73	69.50	72.02	96.40
cycling	197	7	3	2793	96.57	98.50	97.52	99.67
dancing	168	29	32	2771	85.28	84.00	84.63	97.97
drinking	162	32	38	2768	83.51	81.00	82.23	97.67
eating	167	21	33	2779	88.83	83.50	86.08	98.20
fighting	156	22	44	2778	87.64	78.00	82.54	97.80
hugging	173	39	27	2761	81.60	86.50	83.98	97.80
laughing	144	41	56	2759	77.84	72.00	74.81	96.77
music	143	102	57	2698	58.37	71.50	64.27	94.70
running	178	34	22	2766	83.96	89.00	86.41	98.13
sitting	159	149	41	2651	51.62	79.50	62.60	93.67
sleeping	160	11	40	2789	93.57	80.00	86.25	98.30
texting	137	46	63	2754	74.86	68.50	71.54	96.37
using_laptop	138	26	62	2774	84.15	69.00	75.82	97.07

Table 2. Performance of ViT large models

Class	TP	FP	FN	TN	Precision	Recall	F1-score	Accuracy
calling	132	40	68	2760	76.74	66.00	70.97	96.40
clapping	145	41	55	2759	77.96	72.50	75.13	96.80
cycling	198	1	2	2799	99.50	99.00	99.25	99.90
dancing	162	34	38	2766	82.65	81.00	81.82	97.60
drinking	173	22	27	2778	88.72	86.50	87.59	98.37
eating	178	37	22	2763	82.79	89.00	85.78	98.03
fighting	180	62	20	2738	74.38	90.00	81.45	97.27
hugging	174	23	26	2777	88.32	87.00	87.66	98.37
laughing	152	41	48	2759	78.76	76.00	77.35	97.03
music	154	87	46	2713	63.90	77.00	69.84	95.57
running	166	10	34	2790	94.32	83.00	88.30	98.53
sitting	124	51	76	2749	70.86	62.00	66.13	95.77
sleeping	167	25	33	2775	86.98	83.50	85.20	98.07
texting	145	43	55	2757	77.13	72.50	74.74	96.73
using_laptop	173	60	27	2740	74.25	86.50	79.91	97.10

From Table 5 we can see that the SwinT large achieves some of the best individual-model performance with very high precision and recall on most activities. Cycling, for instance, achieves 97.56% precision with perfect recall (100%), and eating achieves 93.19% precision with 89% recall. Larger model size does improve learning of fine features seen in improved performance on calling (75.26% precision) and fighting (85.51% precision, 88.5% recall). Some.classes like clapping continue to demonstrate precision at 66.41% with high recall, which suggests a false positive bias.

From Table 6 we can see that the SwinT small, while improved relative to ViT Small, lags larger models in precision and recall classifying for most classes. Precision in evoking (55.56%) and clapping (57.74%) is low, reflecting challenges with subtle action discrimination. Yet, cycling (97.04% precision) and eating (91.81% precision) are accurately identified. Low recall in texting (55%) and laughing (66.5%) indicates some prediction loss. This model might be a good fit for resource-limited settings but compromises precision on highly complex or visually uncertain actions, illustrating the compromise between performance and model size.

Table 3. Performance of ViT small models

Class	TP	FP	FN	TN	Precision	Recall	F1-score	Accuracy
calling	136	84	64	2716	61.82	68.00	64.76	95.07
clapping	133	71	67	2729	65.20	66.50	65.84	95.40
cycling	194	3	6	2797	98.48	97.00	97.73	99.70
dancing	154	43	46	2757	78.17	77.00	77.58	97.03
drinking	150	12	50	2788	92.59	75.00	82.87	97.93
eating	166	9	34	2791	94.86	83.00	88.53	98.57
fighting	181	126	19	2674	58.96	90.50	71.40	95.17
hugging	181	54	19	2746	77.02	90.50	83.22	97.57
laughing	159	115	41	2685	58.03	79.50	67.09	94.80
music	129	65	71	2735	66.49	64.50	65.48	95.47
running	151	15	49	2785	90.96	75.50	82.51	97.87
sitting	118	54	82	2746	68.60	59.00	63.44	95.47
sleeping	129	16	71	2784	88.97	64.50	74.78	97.10
texting	122	45	78	2755	73.05	61.00	66.49	95.90
using_laptop	143	42	57	2758	77.30	71.50	74.29	96.70

Table 4. Performance of SwinT base models

Class	TP	FP	FN	TN	Precision	Recall	F1-score	Accuracy
calling	136	68	64	2732	66.67	68.00	67.33	95.60
clapping	164	67	36	2733	71.00	82.00	76.10	96.57
cycling	199	2	1	2798	99.00	99.50	99.25	99.90
dancing	165	32	35	2768	83.76	82.50	83.12	97.77
drinking	168	13	32	2787	92.82	84.00	88.19	98.50
eating	179	17	21	2783	91.33	89.50	90.40	98.73
fighting	171	34	29	2766	83.41	85.50	84.44	97.90
hugging	169	17	31	2783	90.86	84.50	87.56	98.40
laughing	145	41	55	2759	77.96	72.50	75.13	96.80
music	140	50	60	2750	73.68	70.00	71.79	96.33
running	182	39	18	2761	82.35	91.00	86.46	98.10
sitting	139	53	61	2747	72.40	69.50	70.92	96.20
sleeping	153	22	47	2778	87.43	76.50	81.60	97.70
texting	131	58	69	2742	69.31	65.50	67.35	95.77
using_laptop	171	75	29	2725	69.51	85.50	76.68	96.53

Table 5. Performance of SwinT large models

Class	TP	FP	FN	TN	Precision	Recall	F1-score	Accuracy
calling	146	48	54	2752	75.26	73.00	74.11	96.60
clapping	170	86	30	2714	66.41	85.00	74.56	96.13
cycling	200	5	0	2795	97.56	100.00	98.77	99.83
dancing	174	42	26	2758	80.56	87.00	83.65	97.73
drinking	181	27	19	2773	87.02	90.50	88.73	98.47
eating	178	13	22	2787	93.19	89.00	91.05	98.83
fighting	177	30	23	2770	85.51	88.50	86.98	98.23
hugging	171	11	29	2789	93.96	85.50	89.53	98.67
laughing	157	44	43	2756	78.11	78.50	78.30	97.10
music	139	42	61	2758	76.80	69.50	72.97	96.57
running	170	22	30	2778	88.54	85.00	86.73	98.27
sitting	141	80	59	2720	63.80	70.50	66.98	95.37
sleeping	150	20	50	2780	88.24	75.00	81.08	97.67
texting	128	29	72	2771	81.53	64.00	71.71	96.63
using_laptop	164	55	36	2745	74.89	82.00	78.28	96.97

Table 6. Performance of SwinT small model

Class	TP	FP	FN	TN	Precision	Recall	F1-score	Accuracy
calling	125	100	75	2700	55.56	62.50	58.82	94.17
clapping	138	101	62	2699	57.74	69.00	62.87	94.57
cycling	197	6	3	2794	97.04	98.50	97.77	99.70
dancing	158	55	42	2745	74.18	79.00	76.51	96.77
drinking	145	19	55	2781	88.41	72.50	79.67	97.53
eating	157	14	43	2786	91.81	78.50	84.64	98.10
fighting	164	52	36	2748	75.93	82.00	78.85	97.07
hugging	159	93	41	2707	63.10	79.50	70.35	95.53
laughing	133	55	67	2745	70.74	66.50	68.56	95.93
music	120	63	80	2737	65.57	60.00	62.66	95.23
running	156	30	44	2770	83.87	78.00	80.83	97.53
sitting	122	92	78	2708	57.01	61.00	58.94	94.33
sleeping	139	26	61	2774	84.24	69.50	76.16	97.10
texting	110	64	90	2736	63.22	55.00	58.82	94.87
using_laptop	151	56	49	2744	72.95	75.50	74.20	96.50

From Table 7 we can see that the ensemble model proposed clearly outperforms every single model individually, with near-perfect precision and recall on almost all classes. Clapping and drinking, for instance, both score 100% precision and recall, illustrating perfect classification. Even the most challenging tasks like texting and laptop are significantly enhanced with precision above 87% and recall above 81%. The ensemble approach reduces false positives and false negatives considerably, resulting in F1-scores above 90% in almost all classes and accuracy above 97%. This confirms that the combination of differing model strengths is capable of successfully counteracting individual flaws to yield strong and stable human activity recognition performance suitable for safety-critical real-world applications.

Table 7. Performance of proposed ensemble model

Class	TP	FP	FN	TN	Precision	Recall	F1-score	Accuracy
calling	61	4	5	920	93.85	92.40	93.13	99.09
clapping	66	0	0	924	100.00	100.00	100.00	100.00
cycling	64	2	2	922	96.97	97.00	96.97	99.60
dancing	62	4	4	920	93.94	93.90	93.94	99.19
drinking	65	0	1	924	100.00	98.50	99.24	99.90
eating	59	3	7	921	95.16	89.40	92.19	98.99
fighting	66	3	0	921	95.65	100.00	97.78	99.70
hugging	64	2	2	922	96.97	97.00	96.97	99.60
laughing	64	2	2	922	96.97	97.00	96.97	99.60
music	64	4	2	920	94.12	97.00	95.52	99.39
running	64	2	2	922	96.97	97.00	96.97	99.60
sitting	62	3	4	921	95.38	93.90	94.66	99.29
sleeping	63	7	3	917	90.00	95.50	92.65	98.99
texting	61	7	5	917	89.71	92.40	91.04	98.79
using_laptop	54	8	12	916	87.10	81.80	84.38	97.98

4. CONCLUSION

We proposed a novel way of performing HAR with static images leveraging the complementing strength of SwinT and ViT in an ensemble architecture. Pondering over the fact that different transformer variants can understand unique spatial and contextual knowledge, we combined six models: swin small, base, large, and ViT small, base, large, in a stacking ensemble framework. This allowed the model to enhance handling of inherent complexity and diversity of human activity in still images. The ensemble ran strongly across activity classes, effectively canceling out weaknesses of individual models. Under thorough evaluation, we probed model behavior, optimization techniques, activation functions, and mis-classification trends. Mis-classifications were more frequent among visually ambiguous classes like sitting, dancing, calling, and use of a laptop, indicating trouble with single-label classification. While the ensemble worked well, it still had a hard time distinguishing overlapping or visually confounded activities since there was no temporal context and single-label classification has a limitation. Computational expense of the ensemble could also deter deployment in resource-scarce or real-time platforms. Upcoming research will study video-based datasets to incorporate motion dynamics, apply multi-label classification to better capture overlapping real-world activity and use attention-based fusion to enhance feature discrimination. Ensemble optimization for efficiency and data set size expansion to cover a wider set of environments and activities will also be extended to enlarge generalizability and practical application in areas like smart surveillance, healthcare monitoring, and human-computer interaction.

FUNDING INFORMATION

The authors state that no funding was involved in supporting this research work.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Rezwana Karim	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Afsana Begum		✓				✓		✓	✓		✓	✓		
Miskatul Jannat	✓		✓	✓			✓			✓	✓		✓	✓
Abu Kowshir Bitto		✓			✓		✓			✓		✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The dataset used in this study is publicly on Kaggle, available at: https://www.kaggle.com/datasets/meetnagadia/human-action-recognition-har-dataset

REFERENCES

[1] C. Dhiman and D. K. Vishwakarma, “A review of state-of-the-art techniques for abnormal human activity recognition,” *Engineering Applications of Artificial Intelligence*, vol. 77, pp. 21–45, Jan. 2019, doi: 10.1016/j.engappai.2018.08.014.

[2] F. Kulsoom, S. Narejo, Z. Mehmood, H. N. Chaudhry, A. Butt, and A. K. Bashir, “A review of machine learning-based human activity recognition for diverse applications,” *Neural Computing and Applications*, vol. 34, no. 21, pp. 18289–18324, 2022, doi: 10.1007/s00521-022-07665-9.

[3] M. A. Khan, M. Mittal, L. M. Goyal, and S. Roy, “A deep survey on supervised learning based human detection and activity classification methods,” *Multimedia Tools and Applications*, vol. 80, no. 18, pp. 27867–27923, 2021, doi: 10.1007/s11042-021-10811-5.

[4] A. K. Bitto and I. Mahmud, “Multi categorical of common eye disease detect using convolutional neural network: a transfer learning approach,” *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 4, pp. 2378–2387, Aug. 2022, doi: 10.11591/eei.v11i4.3834.

[5] J. Yu *et al.*, “Ensemble early exit network on human activity recognition using wearable sensors,” *Computer Networks*, vol. 269, 2025, doi: 10.1016/j.comnet.2025.111409.

[6] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in vision: a survey,” *ACM Computing Surveys*, vol. 54, no. 10, 2022, doi: 10.1145/3505244.

[7] K. Han *et al.*, “A survey on vision transformer,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023, doi: 10.1109/TPAMI.2022.3152247.

[8] S. Huan *et al.*, “A lightweight hybrid vision transformer network for radar-based human activity recognition,” *Scientific Reports*, vol. 13, no. 1, pp. 1–12, 2023, doi: 10.1038/s41598-023-45149-5.

[9] H. Ullah and A. Munir, “Human activity recognition using cascaded dual attention CNN and bi-directional GRU framework,” *Journal of Imaging*, vol. 9, no. 7, 2023, doi: 10.3390/jimaging9070130.

[10] R. K. Vaghela, J. A. Patel, and K. Modi, “Human activity recognition using feature fusion,” *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology*, vol. 14, no. 2, pp. 288–293, 2022, doi: 10.18090/samriddhi.v14spli02.25.

[11] M. G. Morshed, T. Sultana, A. Alam, and Y. K. Lee, “Human action recognition: a taxonomy-based survey, updates, and opportunities,” *Sensors*, vol. 23, no. 4, 2023, doi: 10.3390/s23042182.

[12] K. Alomar, H. I. Aysel, and X. Cai, “CNNs, RNNs and transformers in human action recognition: a survey and a hybrid model,” *Artificial Intelligence Review*, vol. 58, 2024, doi: 10.1007/s10462-025-11388-3.

[13] A. Ulhaq, N. Akhtar, G. Pogrebna, and A. Mian, “Vision transformers for action recognition: a survey.” *arXiv:2209.05700*, 2022.

[14] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: transformers for image recognition at scale,” in *ICLR 2021 - 9th International Conference on Learning Representations*, 2021, pp. 1–22.

[15] Z. Liu *et al.*, “Swin transformer: hierarchical vision transformer using shifted windows,” *2021 IEEE/CVF International Conference on Computer Vision (ICCV),* Montreal, QC, Canada, Feb. 2021, pp. 9992-10002, doi: 10.1109/ICCV48922.2021.00986.

[16] J. Wensel, H. Ullah, and A. Munir, “ViT-ReT: vision and recurrent transformer neural networks for human activity recognition in videos,” *IEEE Access*, vol. 11, pp. 72227–72249, 2023, doi: 10.1109/ACCESS.2023.3293813.

[17] D. R. Reda, F. Chaieb, H. Drira, and A. Aberkane, “ConViViT-a deep neural network combining convolutions and factorized self-attention for human activity recognition,” in *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP),* Poitiers, France, 2023, pp. 1-6, doi: 10.1109/MMSP59012.2023.10337696.

[18] H. Han, H. Zeng, L. Kuang, X. Han, and H. Xue, “A human activity recognition method based on vision transformer,” *Scientific Reports*, vol. 14, no. 1, Jul. 2024, doi: 10.1038/s41598-024-65850-3.

[19] T. Wang, G. Zhou, Y. Pu, R. Moreno, and G. Yang, "Gait recognition with global–local feature fusion based on swin transformer-3DCNN," *Signal, Image and Video Processing*, vol. 19, no. 1, pp. 1–9, 2025, doi: 10.1007/s11760-024-03612-4.

[20] Y. Djenouri and A. N. Belbachir, "A hybrid visual transformer for efficient deep human activity recognition," in *2023 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2023*, 2023, pp. 721–730, doi: 10.1109/ICCVW60793.2023.00080.

[21] D. R. Rani and C. J. Prabhakar, "Vision transformer-based model for human action recognition in still images," *Journal of Computational Analysis and Applications*, vol. 33, no. 8. pp. 522–531, 2024.

[22] L. Nanni, A. Lumini, and C. Fantozzi, "Exploring the potential of ensembles of deep learning networks for image segmentation," *Information*, vol. 14, no. 12, 2023, doi: 10.3390/info14120657.

[23] Z. Zhong *et al.*, "Estimation of bus passenger attributes using swin transformer," in *ACM International Conference Proceeding Series*, Sep. 2022, pp. 121–128, doi: 10.1145/3573942.3573961.

[24] M. Nagadia, "Human action recognition (HAR) dataset," *Kaggle*. 2025. Accessed: May 22, 2025. [Online]. Available: https://www.kaggle.com/datasets/meetnagadia/human-action-recognition-har-dataset

[25] M. Jannat, R. Karim, N. Z. Islam, A. N. Chy, and A. K. M. Masum, "Human activity recognition using ensemble of CNN-based transfer learning models," in *2023 IEEE International Conference on Computing, ICOCO 2023*, 2023, pp. 112–117, doi: 10.1109/ICOCO59262.2023.10398022.

BIOGRAPHIES OF AUTHORS

Rezwana Karim earned her B.Sc. in Computer Science and Engineering from the International Islamic University Chittagong (IIUC) and is currently pursuing an M.Sc. in Software Engineering (Data Science) at Daffodil International University. Her research interests include machine learning, computer vision, and AI applications in healthcare and agriculture. With experience in Python, PyTorch, and TensorFlow, she focuses on developing intelligent systems for disease detection using image data. She can be contacted at email: rezwanaiiuc@gmail.com.

Afsana Begum is an Assistant Professor in the Department of Software Engineering at Daffodil International University, Bangladesh, and is pursuing her Ph.D. at Universiti Malaysia Perlis. She completed her M.Sc. in IIT from the University of Dhaka (1st position) and her B.Sc. from Hajee Mohammad Danesh Science and Technology University (4th position). Her research interests include data science, machine learning, networking, and cybersecurity. She can be contacted at email: afsana.swe@diu.edu.bd.

Miskatul Jannat is currently a Lecturer in the Department of Computer Science and Engineering at the International Islamic University Chittagong (IIUC), Bangladesh, and is pursuing her M.Sc. in the same field. She previously served as faculty at Daffodil International University (2024-2025). Her research interests include machine learning, data science, and large language models (LLMs), with a focus on AI applications in natural language processing and predictive analytics. She can be contacted at email: miskat@iiuc.ac.bd.

Abu Kowshir Bitto is currently working as an AI Solution Specialist at the BRAC which is worlds largest NGO. Previously he worked as Data Scientist at the Centre for Data Science and Research, where he has led and contributed to several impactful initiatives, including projects funded by the Government of Bangladesh and UNESCO. Previously, he served as a Research and Development Engineer at MediprospectsAI Limited, where he led a prestigious Innovate UK-funded research project. He holds both a Bachelor of Science (B.Sc.) and a Master of Science (M.Sc.) degree in Software Engineering with a major in Data Science from Daffodil International University (DIU), Dhaka, Bangladesh. His research affiliations include the Computational Intelligence Lab at Southeast University, the Data Science Lab at DIU, and the Virtual Multidisciplinary Research Lab. He serves as a sessional reviewer for several Scopus-indexed journals and has published multiple papers in Scopus and Web of Science-indexed journals and conferences. His primary research interest is in computer vision. He can be contacted at email: abu.kowshir777@gmail.com.