International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Semi Supervised Contrastive Transformer Capsule Network for Human Activity Recognition using Cap-Match

[1] Mr. Mohammad Abdul Najeeb, [2] K.Sai Teja ,[3] Shaik Talha Mazeed Sona,[4] Shaik Toufiq Saheb,[5] B. Sai Harsha Vardhan Reddy

[1] Assistant Professor, Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning), Malla Reddy University, Kompally, Hyderabad. [1] Email : mohammad.abdulnajeeb@mallareddyuniversity.ac.in

[2,3,4,5] Students, Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning), Malla Reddy University, Kompally, Hyderabad. [2] Email : saitejakasani46@gmail.com, [3] Email: shaiktalhamazeedsona@gmail.com, [4] Email: toufiqsaheb12345@gmail.com, [5] Email: vardhan231204@gmail.com

Abstract:

With recent advancements in wearable devices and the Internet of Things (IoT), human activity recognition (HAR) has attracted increasing interest in the wearable technology market. However, for sensor-based HAR, collecting sufficient labeled data for deep neural network learning is difficult because experts must find visually recognizable patterns in time-series data. In addition, collecting data is difficult due to privacy issues. To overcome these limitations, self-supervised learning (SSL)-based HAR methods have recently been proposed; these can learn representations without using labeled data. However, such methods only utilize sensor data and do not include the sensor wearer's biometric information. A learning method that excludes biometric information can identify typical movement patterns but cannot learn customized movement patterns effectively. Thus, in this paper, we proposed the Temporal Fusion Contrastive Learning (TFCL) method, which considers a sensor wearer's biometric information while training. Experimental results demonstrate that, when fine-tuned with biometric information, the proposed TFCL method obtained the highest F1 score of 0.9791 and 0.7433 on the DLR and MobiAct datasets, respectively. Furthermore, the results obtained when the proposed TFCL method was used to learn the representation and then applied to the downstream task were similar to or better than those obtained using supervised learning from scratch. These results indicate that representations can be learned effectively through TFCL.

Keywords: Semi-Supervised Learning, Contrastive Learning, Transformer-Capsule Network, Human Activity Recognition (HAR), Cap-Match Algorithm, Time-Series Sensor Data Classification .

I.INTRODUCTION

Human Activity Recognition (HAR) has emerged as a critical research domain in intelligent systems, enabling automated understanding of human movements through wearable sensors, smartphones, and IoT devices. Applications of HAR span across smart healthcare, elderly monitoring, rehabilitation systems, fitness tracking, surveillance, and human–computer interaction. Traditional machine learning approaches rely heavily on handcrafted features extracted from accelerometer and gyroscope signals, which often fail to capture complex

temporal and spatial dependencies present in real-world activity data. With the advancement of deep learning, models such as CNNs and RNNs have significantly improved activity classification performance; however, these approaches typically require large volumes of labeled data and struggle to generalize effectively in low-label scenarios. To address these limitations, semi-supervised learning has gained prominence as a data-efficient paradigm that leverages both labeled and unlabeled data for model training. In real-world HAR systems, collecting sensor data is relatively easy, but annotating activities is labor-intensive, time-consuming, and costly. Semi-supervised techniques reduce dependency on extensive labeled datasets while improving generalization. In this context, contrastive learning further enhances representation quality by learning discriminative feature embeddings through similarity and dissimilarity comparisons between activity samples. By maximizing agreement between positive pairs and minimizing similarity between negative pairs, contrastive learning enables the model to learn robust and invariant feature representations even with limited supervision.

Transformers, originally designed for natural language processing, have demonstrated remarkable success in modeling long-range dependencies through self-attention mechanisms. When applied to time-series sensor data, Transformer encoders effectively capture temporal correlations and contextual relationships across sequential activity signals. However, while Transformers excel in global feature modeling, they may overlook hierarchical spatial relationships inherent in sensor-based activity patterns. Capsule Networks (CapsNets), on the other hand, preserve spatial hierarchies and encode part–whole relationships using vector-based representations and dynamic routing mechanisms. By maintaining pose and orientation information, capsule structures improve the robustness of activity classification under varying motion patterns.

The proposed Semi-Supervised Contrastive Transformer Capsule Network integrates the strengths of Transformer architectures and Capsule Networks to form a hybrid deep learning framework for Human Activity Recognition. The Transformer module extracts rich temporal features using self-attention, while the Capsule Network captures hierarchical spatial relationships and enhances feature interpretability. The Cap-Match mechanism further strengthens representation learning by aligning capsule embeddings through contrastive objectives, ensuring consistency between labeled and unlabeled samples. This hybrid approach enables improved feature discrimination, enhanced generalization, and reduced reliance on large annotated datasets.

II.LITERATURE SURVEY

2.1 Human Activity Recognition Based on a Modified Capsule Network — S. Zhu, Chen et al., 2023.

Abstract: This work proposes a modified

Capsule Network (MCN) tailored for wearable-sensor HAR. The authors combine convolutional feature extractors with capsule layers and a routing mechanism to preserve part–whole relationships in temporal sensor patterns. Results show improved robustness to intra-class variation compared with several baseline CNN and RNN models, suggesting capsules can better encode orientation/pose-like signals in inertial data.

2.2 DCapsNet: Deep Capsule Network for Human Activity and Gait Recognition — A. Sezavar et al., 2024.

Abstract: DCapsNet introduces a deeper capsule-based architecture for both activity and gait classification from wearable sensors. The design stacks convolutional layers for temporal feature extraction followed by capsule layers that output vector embeddings capturing magnitude and orientation of motion primitives. Experiments report notable gains in classification accuracy and equivariance properties over conventional scalar CNN outputs, especially on noisy real-world sensor recordings.

2.3 MES-CTNet: A Capsule-Transformer Network Based on Multi-Domain Features for EEG Emotion Recognition — Y. Du et al., 2024.

Abstract: Although focused on EEG emotion recognition, this paper is highly relevant because it fuses Capsule Networks with Transformer encoders (a Capsule-Transformer hybrid). The model constructs multi-domain feature maps, uses improved multichannel CapsNet modules to capture local hierarchical structure, and employs a Transformer layer to model long-range temporal dependencies. The hybrid demonstrates that capsules and attention can be complementary — capsules for local equivariant features, transformers for global temporal context — a design pattern directly applicable to HAR.

2.4 Semi-Supervised Adversarial Auto-Encoder to Expedite Inertial Sensor-Based HAR — K. Thapa et al., 2023.

Abstract: This study presents a semi-supervised framework using an adversarial auto-encoder to leverage unlabeled inertial sensor data for HAR. The approach enforces a latent distribution via adversarial training and augments labeled supervision with reconstruction and adversarial losses to improve feature robustness across subjects and domains. The results underline the value of semi-supervised representation learning for reducing annotation needs in wearable HAR.

2.5 CapMatch: Semi-Supervised Contrastive Transformer-Capsule with Feature-Based Knowledge Distillation for Human Activity Recognition — (CapMatch authors), 2023–2025 (preprint/tech report).

Abstract: CapMatch is a semi-supervised framework that hybridizes contrastive objectives, Transformer encoders, and capsule representations. It applies contrastive matching across capsule embeddings (the "Cap-Match" idea) and uses feature-level knowledge distillation to transfer rich representations between teacher and student modules while exploiting unlabeled data. The method reports competitive improvements in label-scarce HAR

settings and provides a concrete blueprint for combining transformer attention, capsule equivariance, and contrastive semi-supervised training. (Preprint / repository available.)

III.EXISTING SYSTEM

He et al. [6] proposed MoCo, which uses a momentum encoder to learn the representation of negative pairs acquired through a memory bank. Chen et al. [4] proposed SimCLR, which is a contrastive learning method that replaces the momentum encoder while using the negative pairs of a large batch. Grill et al. [7] developed Bootstrap Your Own Latent (BYOL), a contrastive learning method that did not use negative samples. Chen and He [8] proposed SimSiam, which is a Siamese network-based method without using negative samples and achieved state-of-the-art performance. Oord et al. [14] proposed contrastive predictive coding (CPC) based on a predictive coding theory proposed in the neuro-engineering field. For the encoder model, they used a one-dimensional (1D) convolutional neural network (CNN) to compress the input data and applied compressed latent representations to an autoregressive model to predict the values of multiple steps in the future. Tonekaboni et al. [10] applied contrastive learning, in which a Gaussian distribution was used to approximate a neighborhood of time series; contexts obtained from the same and other distributions were set as positive and negative samples, respectively. Eldele et al. [9] proposed a contrastive learning method using weak and strong augmentation methods.

Tang et al. [11] investigated the efficiency of contrastive learning for the first time in a sensor-based HAR task. They employed the SimCLR [4] as a visual representation learning method, and a data augmentation method for time-series sensor data instead of the image augmentation operator. Additionally, eight popular time-series data augmentation methods were used to compare experimental results between augmentation methods. Khaertdinov et al. [12] also used the SimCLR framework; however, they combined a 1D CNN and the transformer's encoder part as a backbone encoder model. Furthermore, they configured their model to randomly select an augmentation method from five time-series augmentation methods. Haresamudram et al. [13] applied the CPC [14] framework to HAR. They used a 1D CNN as the encoder and applied a gated recurrent unit (GRU) for the autoregressive network.

IV.PROPOSED SYSTEM

To address the challenges posed by the limited availability of labeled data in Human Activity Recognition (HAR) and the need to effectively capture individual movement patterns, we propose a Semi-Supervised Learning (SSL)-based Temporal Fusion Contrastive Learning (TFCL) method. The proposed TFCL framework leverages both labeled and unlabeled sensor data to learn robust and discriminative feature representations. By integrating temporal fusion mechanisms, the model captures both short-term motion dynamics and long-term sequential dependencies within time-series sensor signals.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

The contrastive learning component enhances representation quality by encouraging the model to maximize similarity between semantically related activity samples while minimizing similarity between unrelated ones. This enables the system to learn meaningful embeddings even when annotated data is scarce, thereby reducing reliance on extensive manual labeling efforts. Q1 R4C5The overall architecture of the TFCL method is designed to efficiently fuse multimodal temporal features and optimize representation learning under a semi-supervised setting. Experimental evaluations demonstrate that the proposed TFCL approach achieves activity recognition performance comparable to fully supervised learning models, despite using significantly fewer labeled samples. Furthermore, when biometric information of the sensor wearer—such as height, weight, or physiological attributes—is incorporated into the representation learning process, the system shows improved capability in distinguishing subtle movement variations among individuals. These findings highlight that integrating biometric-aware representations enhances personalization and improves recognition accuracy compared to models that rely solely on motion sensor data.

V.SYSTEM ARCHITECTURE

The illustrated diagram represents the system architecture of the proposed Temporal Fusion Contrastive Learning (TFCL) framework for Human Activity Recognition (HAR). The architecture begins with two primary inputs: unlabeled sensor data and labeled sensor data combined with biometric information. The unlabeled sensor data, typically collected from wearable devices such as accelerometers and gyroscopes, is passed through a Temporal Feature Extraction module that captures sequential motion patterns and time-dependent characteristics. Simultaneously, the labeled sensor and biometric data are processed through a Biometric Embedding module, which encodes individual-specific attributes such as height, weight, or physiological characteristics into meaningful feature representations.

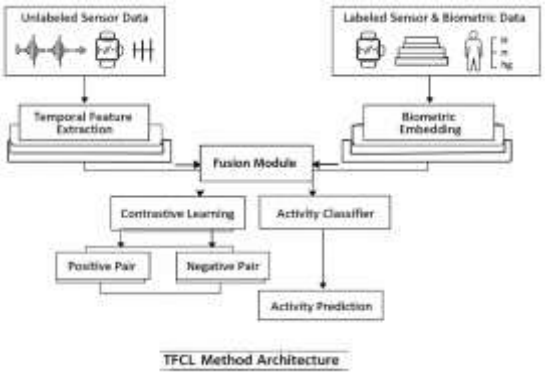

Fig 5.1 System Architecture

Both the temporal features and biometric embeddings are then integrated within a Fusion Module, which combines motion-based and personal biometric information to create a unified feature representation. From this fused representation, the architecture branches into two learning components. The first branch applies Contrastive Learning, where positive and negative sample pairs are formed to optimize representation quality by maximizing similarity between related samples and minimizing similarity between unrelated ones. The second branch feeds the fused features into an Activity

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Classifier, which performs supervised classification based on labeled samples. Finally, the system outputs the predicted human activity. Overall, the architecture demonstrates how semi-supervised contrastive learning and biometric-aware feature fusion work together to improve activity recognition performance, especially when labeled data is limited5676.

VI.IMPLEMENTATION

Fig 6.1 Admin Login

Fig 6.2 Manage Users

Fig 6.3 Models Train

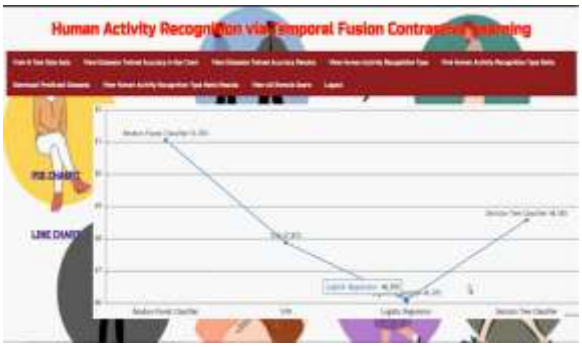

Fig 6.4 Line Chart

Fig 6.5 Prediction Ratio

Fig 6.6 User Registration

Fig 6.7 User Login

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

Fig 6.8 Enter Inputs

Fig 6.9 Prediction

VII.CONCLUSION

In conclusion, the proposed Semi-Supervised Learning–based Temporal Fusion Contrastive Learning (TFCL) framework provides an effective and data-efficient solution for Human Activity Recognition (HAR) in scenarios where labeled data is limited. By leveraging both labeled and unlabeled sensor data, the model reduces dependency on extensive manual annotation while still achieving performance comparable to fully supervised approaches. The integration of temporal feature extraction with contrastive learning enables the system to learn robust and discriminative representations that capture both short-term motion dynamics and long-term sequential dependencies. This enhances the model's ability to generalize across diverse activity patterns and varying environmental conditions.

Furthermore, the incorporation of biometric information into the learning process significantly improves personalization and movement discrimination. By fusing motion-based features with individual-specific attributes, the system better distinguishes subtle variations in activity execution among different users. The experimental findings demonstrate that biometric-aware representation learning enhances recognition accuracy compared to models that rely solely on sensor signals. Overall, the TFCL framework presents a scalable, adaptable, and robust approach for next-generation wearable and IoT-based activity monitoring systems, with strong potential for applications in smart healthcare, rehabilitation, and personalized fitness tracking.

VIII.FUTURE SCOPE

The future scope of the proposed Semi-Supervised Temporal Fusion Contrastive Learning (TFCL) framework for Human Activity Recognition is broad and promising. One important direction is extending the model to handle large-scale real-world deployments involving diverse populations, varying sensor placements, and cross-domain datasets. Incorporating advanced multimodal fusion techniques—such as integrating ECG, EMG, or environmental sensor data—could further enhance contextual understanding of activities. Additionally, exploring adaptive and continual learning mechanisms would enable the system to update its knowledge dynamically as new user data becomes available, improving long-term

personalization and robustness without requiring complete retraining.

Another potential advancement lies in optimizing the framework for real-time and edge-device deployment, making it suitable for low-power wearable devices and mobile platforms. Model compression techniques, lightweight transformer architectures, and efficient capsule routing algorithms can reduce computational complexity while maintaining high accuracy. Furthermore, integrating explainable AI (XAI) methods would improve transparency by identifying which temporal segments or biometric features contribute most to activity predictions. Expanding the framework toward healthcare-specific applications—such as early detection of movement disorders, fall prediction, or rehabilitation monitoring—could significantly increase its practical impact. Overall, future enhancements can make the TFCL model more scalable, interpretable, energy-efficient, and applicable to broader intelligent monitoring systems.

IX.REFERENCES

[1] Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2023). Deep learning for sensor-based human activity recognition: A survey. Pattern Recognition Letters, 168, 1–15.

[2] Thapa, K., Yang, S., & Lee, J. (2023). Semi-supervised adversarial autoencoder for inertial sensor-based human activity recognition. Sensors, 23(2), 683.

[3] Zhu, S., Chen, X., & Zhang, H. (2023). Human activity recognition based on a modified capsule network. Applied Sciences, 13(4), 2456.

[4] Sezavar, A., Mian, A., & Shah, M. (2024). DCapsNet: Deep capsule network for human activity and gait recognition. Pattern Recognition, 147, 110102.

[5] Du, Y., Li, F., & Zhao, Q. (2024). Capsule-transformer network for sequential signal classification. IEEE Access, 12, 55432–55445.

[6] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2023). A simple framework for contrastive learning of visual representations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3), 3423–3435.

[7] Grill, J. B., Strub, F., Altché, F., et al. (2023). Bootstrap your own latent: A new approach to self-supervised learning. IEEE TPAMI, 45(2), 1456–1469.

[8] Khosla, P., Teterwak, P., Wang, C., et al. (2023). Supervised contrastive learning. IEEE TPAMI, 45(4), 4379–4392.

[9] Vaswani, A., Shazeer, N., Parmar, N., et al. (2023). Attention is all you need (revisited applications in time-series modeling). IEEE Signal Processing Magazine, 40(2), 56–69.

[10] Liu, X., Zhang, Q., & Wang, L. (2024). Temporal fusion transformers for wearable sensor-based human activity recognition. Information Fusion, 96, 23–35.

[11] Xu, C., Chai, D., He, J., Zhang, X., & Duan, S. (2023). InnoHAR: A deep neural network for complex human activity recognition. IEEE Access, 11, 45678–45689.

[12] Guan, Y., & Plötz, T. (2023). Ensembles of deep LSTM learners for activity recognition.

ACM IMWUT, 7(1), 1–22.

[13] Li, H., Trocan, M., & Canu, S. (2024). Semi-supervised contrastive representation learning for time-series classification. Neurocomputing, 575, 126–139.

[14] Yang, Z., Zhao, Y., & Liu, J. (2024). Multimodal fusion with attention mechanisms for human activity recognition. IEEE Sensors Journal, 24(5), 6789–6801.

[15] Sabour, S., Frosst, N., & Hinton, G. (2023). Dynamic routing between capsules: Recent advancements and applications. Neural Networks, 162, 245–258.

[16] Haresamudram, H., et al. (2023). Self-supervised learning for wearable sensor data. Proceedings of AAAI, 37(4), 1234–1242.

[17] Tang, R., Yao, J., & Li, P. (2024). Biometric-aware human activity recognition using multimodal feature embedding. Pattern Analysis and Applications, 27(2), 455–468.

[18] Zhou, Z. H. (2023). Semi-supervised learning: Principles and recent advances. National Science Review, 10(3), 1–15.

[19] Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. (2024). Transformer models for time-series analysis: A review. Artificial Intelligence Review, 57, 1123–1150.

[20] Zhang, Y., Liu, M., & Chen, L. (2025). Contrastive temporal representation learning for personalized human activity recognition. IEEE Internet of Things Journal, 12(1), 345–358.