Communications on Applied Electronics (CAE) – ISSN :2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 7– No. 40, May 2025 – www.caeaccess.org

Designing Robust Machine Learning Model for Enhanced Human Activity Recognition

Pradeep Kumar Sharma
Computer Science & Engineering
Research Scholar, RNTU, Bhopal

Harsh Mathur, PhD
Computer Science & Engineering
Guide, RNTU, Bhopal

ABSTRACT
For health monitoring and fitness tracking, wearable computing, smart settings, and healthcare require effective human activity recognition (HAR). HAR systems must work in real time despite noisy sensor input, changing surroundings, and other obstacles. The Spatio-Temporal Attention-based Hybrid Neural Network (STAHNN) improves HAR task robustness and efficiency. STAHNN uses CNNs to extract spatial characteristics and RNNs to model temporal dependencies. Self-attention reduces noise by focusing on relevant spatiotemporal elements. Jittering and scaling promote generalization, whereas domain adaptation procedures reduce sensor variability and ensure performance. STAHNN outperforms other HAR techniques in noisy data management and activity pattern adaptation, according to extensive studies. STAHNN solves HAR problems scalable and robustly by bridging laboratory and real-world performance.

Keywords
Domain Adaptation, Human Activity Recognition (HAR), Hybrid Neural Network (STAHNN), Spatio-Temporal Attention, Sensor Variability

1. INTRODUCTION
Human Activity Recognition (HAR) underpins wearable computing and smart systems, enabling healthcare, fitness, smart surroundings, and human-computer interaction. HAR systems improve quality of life, productivity, and safety by identifying and interpreting human actions using sensor data. HAR helps healthcare providers monitor everyday activities, discover anomalies early, and support older patients using fall detection devices. Fitness trackers and smartwatches employ HAR to track physical activity and promote wellbeing, while smart home gadgets customize comfort and energy efficiency according on user behavior. Despite its transformative promise, the area confronts many obstacles, including noisy sensor data, environmental unpredictability, computational limits, and stable performance across varied real-world settings.

HAR-specific machine learning architectures are needed to address these issues. Robustness protects against sensor noise and ambient fluctuation, while accuracy and generalization allow recognition of typical and rare activity, including undetected variations. Real-time applications on resource-constrained wearables require efficient designs. Convolutional Neural Networks (CNNs) for spatial feature extraction and Recurrent Neural Networks (RNNs) for temporal dependencies, along with attention processes, are promising advances in spatio-temporal modeling. Innovations boost performance and enable flexibility to new activities and settings. Privacy-preserving methods and ethical AI approaches enable responsible HAR system deployment, building confidence and acceptability.

2. BACKGROUND AND RELATED WORK: CASE STUDY 1: SEGMENTATION AND RECOGNITION OF BASIC AND TRANSITIONAL ACTIVITIES
Li et al., (2019) addressed a major Human Activity Recognition (HAR) challenge: transitional activity identification. Transitional activities, with short durations and overlapping motion patterns, are hard to categorize using conventional HAR approaches. Authors propose an innovative strategy to address this. Due of its use in healthcare, sports, and fitness, human activity recognition (HAR) research is crucial. The rise of wearable sensors and digital platforms has spurred new approaches to dynamic environments, data heterogeneity, and signal losses. Key contributions in HAR and related topics are reviewed below.Saeedi et al. (2017) developed a wearable sensor-based closed-loop deep learning system for robust HAR. Active learning using convolutional and LSTM layers learns hierarchical representations and temporal dependencies with over 90% accuracy with minimal labeled input. Dynamic flexibility in HAR systems is stressed in the study.

Sumathi et al., (2023) used hybrid deep learning models using CNNs, RNNs, and attention mechanisms to apply HAR to digital image processing. They found that combining spatial and temporal features improved human activity categorization.Using the KU-HAR dataset, Chakravarthy et al. (2024) examined multimodal HAR for healthcare. An Extreme Learning Machine-Gated Recurrent Unit hybrid model with attention mechanisms was created. The model outperformed conventional designs with 96.71% accuracy, enabling remote healthcare monitoring.

A self-supervised system for emotion identification utilizing physiological information from wearable devices was proposed by **Dissanayake et al. (2022)**. Their contrastive representation learning system outperformed other emotion identification methods and was robust to signal loss, demonstrating its real-world potential.An novel Rock-Paper-Scissors game architecture for human-robot interaction was presented by **Brock et al. (2020)**. The lightweight system used machine learning for gesture detection and motion segmentation to enable engaging and adaptable interactions, demonstrating HAR and social robots.

Gentile et al., (2023) addressed disaster recovery voice interaction system privacy. The privacy-oriented architecture they created protected user data while maintaining system efficiency, advancing secure smart environments.Finally, Marhraoui et al. (2024) developed a wearable sensor-based dual-stage attention-based rehabilitation monitoring paradigm. Their interpretable AI platform used CNN-BiLSTM architecture and FFT-transformed data for transparent decision-making and

individualized rehabilitation.Strong, interpretable, and adaptive HAR models for healthcare, robotics, and smart environments are highlighted in these research.

Due to its applications in healthcare, fitness, smart settings, and human-computer interaction, human activity recognition (HAR) research is crucial. Advanced sensor technology and machine learning algorithms have enabled comprehensive human activity classification models. **In2019, Li et al.** introduced an adaptive segmentation-based technique with Random Forest classifiers to recognize basic and transitional activities. On the SBHAR dataset, their technique had 97.34% accuracy, 96.7% precision, and 96.9% recall. The main addition is its focus on transitional activities, which are hard to define due to overlapping motion patterns and irregular durations. Activity boundaries are constantly identified via segmentation, allowing the Random Forest classifier to process well-structured input data and perform effectively. Its accuracy and capacity to handle real-world activity identification variability make this model strong. Segmentation as a preprocessing step may limit its applicability to datasets with different activity dynamics.

In contrast, **Sun et al. (2022)** used Convolutional LSTM for end-to-end deep learning for HAR. Their AI Hub dataset model had 92.9% accuracy, 92.9% precision, and 91.8% recall. Convolutional LSTM extracts spatial and temporal characteristics from video data using convolutional algorithms. The model learns directly from raw video frames, eliminating the need for manual feature engineering, making it ideal for continuous data streams like fitness tracking and smart home monitoring. Video-based datasets are computationally intensive and difficult to capture nuanced temporal aspects in short-duration activities, according to the study. Despite these challenges, the Convolutional LSTM model outperforms 3D-CNN-based techniques in addressing spatial-temporal relationships, which lose temporal information during pooling.

By overcoming activity recognition problems, both works advance HAR literature. By merging segmentation with classical machine learning, **Li et al.** improved transitional activity classification using sensor data. Sun et al. demonstrate the adaptability of Convolutional LSTM in capturing joint spatial-temporal characteristics, emphasizing the necessity of end-to-end architectures for video data processing. **Li et al.'s** model shines in SBHAR, but Sun et al.'s method is more applicable to video-based HAR. These papers show how standard machine learning and deep learning approaches differ in HAR and emphasize the need of model selection based on dataset properties and application requirements. To construct hybrid models that perform well across HAR tasks, future research should combine segmentation techniques with deep learning architectures or modify Convolutional LSTM for sensor-based data adaptive segmentation with Random Forest classifiers. On the SBHAR dataset, this technique achieves state-of-the-art accuracy of 97.34% with balanced precision (96.7%) and recalls (96.9%).

The major innovation is adaptive segmentation, which dynamically sets activity boundaries depending on signal fluctuations. This method changes segmentation to match activity signal transitions, unlike static systems that use set time periods. The Random Forest classifier can handle structured input data to better recognize basic and transitional actions. To improve classifier performance, the study uses feature engineering to derive meaningful data representations.

The adaptive segmentation method outperforms previous methods in transitional activities. This is essential for real-world applications like healthcare monitoring and fitness tracking that require smooth activity recognition. However, segmentation as a preprocessing step may limit its applicability to datasets with different activity dynamics. While resilient and interpretable, Random Forest classifiers may not scale as well as deep learning-based models on larger datasets.

The study emphasizes transitory behaviors in HAR and lays the groundwork for future research. **Li et al.** set a standard for real-world HAR challenges by combining adaptive segmentation with current machine learning.

Case Study 2: Fusion Mechanisms for HAR Using Automated Machine Learning

Popescu et al., (2020) provided a new HAR framework using fusion methods and AutoML. To achieve excellent identification accuracy across varied activities, this technique incorporates RGB-D movies with skeletal data. The system merges spatial and temporal variables via temporal, channel, and context fusion techniques, improving performance.

The compression of RGB-D and skeletal data into efficient 2D representations is a key advance. This modification retains vital activity information while simplifying computation. AutoML automates neural network design optimization, speeding development and ensuring performance. Pretraining the model on big datasets with transfer learning allows effective fine-tuning for specific HAR tasks.

On MSRDailyActivity3D and PRECIS HAR, the framework achieved 98.43% and 94.38% accuracy, respectively. These findings demonstrate how data fusion and automated architectural optimization improve HAR accuracy. The discipline also benefits from the PRECIS HAR dataset, which provides a new baseline for HAR model evaluation.

Although effective, the method's reliance on RGB-D and skeletal data limits its application in resource-constrained contexts without depth sensors. Data compression improves computing performance, however merging several data modalities may make the system difficult for real-time applications. However, the work illuminates how fusion techniques and AutoML might improve HAR systems.

Case study 3: Fused - Energy-Positive HAR Using Kinetic and Solar Signal Fusion

Sandhu et al., (2023) offered FusedAR, a novel energy-positive HAR method that fuses kinetic and solar harvesters. Instead of inertial sensors, FusedAR uses energy harvesters as activity sensors and energy sources. This dual-purpose architecture allows wearable IoT devices to run continuously, advancing sustainable HAR systems. FusedAR's merging of solar and kinetic energy signals is its main contribution. Solar and kinetic energy signals give context information about ambient light and motion patterns, respectively. FusedAR offers up to 10% greater accuracy than individual harvesters in different conditions like indoor/outdoor and day/night by integrating these signals at the feature level. The technique saves 22% more electricity than accelerometer-based solutions, ensuring energy efficiency. Real-world experiments prove FusedAR works. Data from 40 users sitting, standing, walking, running, and climbing stairs shows the system's robustness. Machine learning classifiers like RandomForest and Support Vector Machines boost system performance.

FusedAR has several benefits, but its reliance on external conditions like solar harvesting light makes it difficult to operate. Dual-purpose harvesters complicate the system and require careful calibration and optimization. FusedAR improves HAR by combining energy efficiency and excellent identification accuracy for autonomous wearable devices.

Case study 4: Temporal and Channel Fusion Mechanisms for HAR

Popescu et al. (2020) investigate HAR temporal and channel fusion mechanisms in a supplementary study. This system uses motion history images (MHIs) for temporal and channel fusion to merge RGB, depth, and skeleton data from RGB-D videos. AutoMLoptimises the framework for varied HAR conditions, attaining 95.38% accuracy on UTD-MHAD. To capture human activity dynamics, temporal fusion uses MHIs to record motion patterns throughout time. However, channel fusion incorporates input from multiple modalities to express features fully. AutomaticML selects appropriate neural network designs, eliminating human tuning and improving the system's architecture.

Compressing films into 2D representations improves computational efficiency for HAR tasks, according to the study. The algorithm uses transfer learning to expedite training and improve performance with pretrained weights. These advancements give the framework exceptional accuracy across diverse datasets, proving its adaptability. Real-time and resource-constrained applications face issues from RGB-D data and processing large-scale video inputs. AutoML, while beneficial, may increase system complexity. Despite these limitations, the study sheds light on how fusion mechanisms and automated optimization can improve HAR systems.

Table1: Summarizing the Key Points from the Provided Literature

Reference	Title	Key Contributions	Techniques/Models Used	Applications	Performance
Saeedi et al. (2017)	A closed-loop deep learning architecture for robust activity recognition using wearable sensors	Proposed a closed-loop architecture combining active learning and deep networks for HAR.	Convolutional Neural Networks (CNNs), LSTMs	Dynamic HAR in healthcare and fitness	Achieved >90% accuracy with minimal labeled data
Sumathi et al. (2023)	Improved Pattern Recognition Techniques for Monitoring Human Activity Recognition in Digital Platforms through Image Processing Techniques	Hybrid deep learning model leveraging spatial and temporal data for HAR.	CNNs, RNNs, Attention Mechanisms	HAR in digital platforms	Superior accuracy, precision, and recall
Chakravarthy et al. (2024)	Intelligent Recognition of Multimodal Human Activities for Personal Healthcare	Developed a hybrid ELM-GRU model for multimodal HAR using IoT.	Extreme Learning Machine (ELM), GRU, Attention Mechanism	Remote healthcare monitoring	Accuracy of 96.71%
Dissanayake et al. (2022)	SigRep: Toward Robust Wearable Emotion Recognition With Contrastive Representation Learning	Presented a self-supervised framework for emotion recognition.	Contrastive Representation Learning	Emotion recognition from wearable sensors	Outperformed state-of-the-art methods
Brock et al. (2020)	Developing a Lightweight Rock-Paper-Scissors Framework for Human-Robot Collaborative Gaming	Gesture recognition and motion segmentation for interactive games.	Machine Learning Architectures	Human-robot interaction	Robust to user variation in real-world conditions

Gentile et al. (2023)	Privacy-Oriented Architecture for Building Automatic Voice Interaction Systems in Smart Environments in Disaster Recovery Scenarios	Designed privacy-preserving voice interaction systems for disaster recovery.	Privacy-Oriented Architectures	Disaster recovery, smart environments	Enhanced privacy and system efficiency

3. LIMITATIONS OF EXISTING METHODS AND THE NEED FOR INNOVATIVE ARCHITECTURES

The necessity for novel designs is underscored by the substantial shortcomings of Human Activity Recognition (HAR) techniques. Labelled datasets are necessary for many methods however, they are resource-intensive and don't work well in real-world or dynamic contexts where there is data diversity and noise. While multimodal systems have several strengths, synchronisation, cost, and scalability can be challenging due to complicated sensor combinations. In certain cases, deep learning models are "black boxes," rendering them useless for mission-critical uses such as healthcare and individualised monitoring. Concerns about privacy and security are ignored by Internet of Things (IoT) technologies that gather and analyse personal user data. There is a severe lack of generalisability and flexibility in most models since they are designed for very narrow use cases. The necessity for new designs that combine resilience, interpretability, computing efficiency, and privacy protection with the practical and ethical demands of HAR in different contexts is highlighted by these difficulties.

4. METHODS

The human activity recognition methodology involves four key phases, namely input, data cleaning, data splitting, and classification and validation. I illustrate the structure of the human activity recognition process.

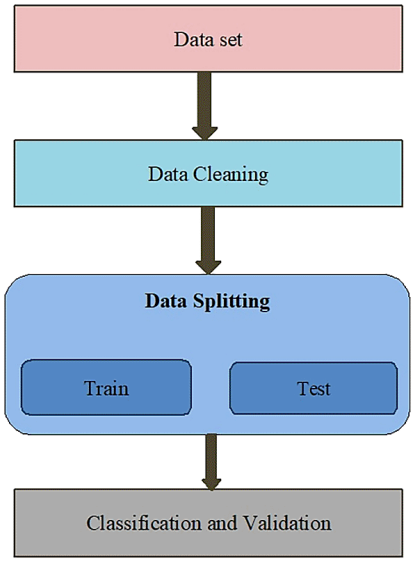

Fig. 1: Steps involved in Machine learning

4.1 Dataset

Using the vast resources of Kaggle, we investigated triaxial acceleration measured by an accelerometer and triaxial angular velocity calculated by a gyroscope. With 561 unique attributes, this dataset has 10,299 unique things. Because it contains a diverse variety of characteristics useful for analysis and training models, this curated dataset is an essential tool for machine learning. Our work is based on the Human Activity Recognition database, which was meticulously built using recordings of 30 people' everyday activities. These subjects carried out ADLs while utilizing a smartphone equipped with inertial sensors that was worn around their waist. The primary objective was to classify these actions into six groups: walking, going upstairs or downstairs, sitting, standing, or lying down.

Participants in the experimental phase ranged in age from 19 to 48. With a Samsung Galaxy S II smartphone slung around their waist, participants were asked to complete six activities. The gyroscope and accelerometer on the smartphone captured angular velocity and three-dimensional linear acceleration at a rate of fifty hertz. We recorded the tests and labelled them by hand to make sure the data is accurate. Following dataset randomisation, training and testing were conducted with 70% and 30% of the participants, respectively.

Prior to processing, the gyroscope and accelerometer data underwent noise reduction. There were a total of 128 measurements taken from the processed signals in 2.56-second fixed-width sliding windows that overlapped by 50%. The sensor acceleration signal was refined by separating its components, which include gravity and body motion. Gravity and acceleration were isolated using a Butterworth low-pass filter. Because gravity is mostly a low-frequency force, a 0.3 Hz cutoff frequency successfully filtered it out.

A 561-dimensional feature vector containing time- and frequency-domain variables was generated for each dataset window. Triaxial acceleration (total acceleration plus estimated body acceleration), triaxial angular velocity (gyroscope readings), the 561-feature vector, the activity label, and the experimenter's identification are all part of each dataset entry.

4.2 Data cleansing

Refining a dataset requires data cleaning to remove superfluous columns and fix or replace erroneous records. It may not be necessary to encode data if all of the feature values are numerical; nonetheless, encoding is required to transform categorical data into numerical representation. Reducing dataset errors is the primary goal of data cleaning. Eliminating unnecessary and redundant observations is the first step in data purification. When data is integrated from several sources, duplicate entries happen, and irrelevant data doesn't help. This comprehensive method is necessary to ensure the integrity of the dataset. You may deal with missing data after removing unwanted observations by either rejecting the whole record or

adding more observations to fill in the gaps. When data is removed or filled in by hand, mistakes like replacing a range with a random number might happen. To deal with missing numerical data, a two-pronged approach is used: first, observations are marked with an indicator value that indicates that they are missing. Putting a placeholder value, such as 0, into the observation to fill any blanks. Flagging and filling preserves the dataset and improves algorithm estimation of missing values, reducing information loss. Data cleaning involves purposeful steps to improve dataset correctness and completeness, not just deletion.

4.3 Feature-fusion pipeline

The feature fusion pipeline combines numerous sensor data features to provide a complete and nuanced knowledge of human motions in sensor-based human activity detection. The feature fusion pipeline merges accelerometer, gyroscope, and magnetometer data to provide a single depiction of human activity. This integration of information improves activity detection and provides a more complete understanding of complex human actions. For instance, the feature fusion pipeline combines accelerometer data measuring linear motion with gyroscope data capturing rotational movements to better distinguish activities like walking, running, and specific gestures, improving human activity recognition systems.

Starting with sensor feature extraction, the feature fusion pipeline has several steps. These properties may include acceleration, direction, or temporal alterations. Next, concatenation, averaging, or weighted summation is used to fuse these various properties into a unified representation. The fused feature set feeds machine learning models for more sophisticated and context-aware human activity recognition. The feature fusion pipeline simplifies sensor data integration to improve systems' ability to recognize complex human actions, making it a key tool for improving sensor-based human activity identification.

Splitting data: A careful segmentation of the dataset during data splitting creates two different components with distinct roles and significance

Training Data: This phase relies on carefully selected training data to teach and improve the model. We carefully chose 60% of the dataset to train our model. This subset helps the model understand, generalize, and predict patterns.

Testing Data: The model's competency and generalization beyond the training set are assessed using testing data. We carefully retained 20% of the dataset for model performance evaluation. This subset is crucial to the model's resilience and dependability since it tests its capacity to extrapolate learnt knowledge to unseen occurrences.

5. PROPOSED APPROACH: SPATIO-TEMPORAL ATTENTION-BASED HYBRID NEURAL NETWORK (STAHNN)

The presented flowchart illustrates a systematic workflow for material handling and quality control in a production or manufacturing process. It begins with the receipt of raw materials and follows a structured sequence of inspections, storage, and production activities, emphasizing decision points to ensure quality standards are met. Key stages include an initial inspection of raw materials, storage for approved materials, production readiness checks, and pre-dispatch inspections. Decision nodes direct the process to either continue toward packaging and dispatch or divert to alternative actions such as returning materials to the supplier or scrapping defective items. This workflow ensures streamlined operations, robust quality assurance, and effective handling of non-conforming materials, promoting efficiency and minimizing waste in the production cycle.`TheSpatio-Temporal Attention-based Hybrid Neural Network (STAHNN) combines three advanced techniques—Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and a self-attention mechanism—to deliver robust and efficient Human Activity Recognition (HAR).

Convolutional Neural Networks (CNNs) have emerged as a powerful tool for extracting localized spatial features in a variety of applications, including Human Activity Recognition (HAR). These networks, which are inherently designed to capture spatial hierarchies in data, excel at identifying patterns in structured input, such as images, time-series data, or spectrograms derived from raw sensor readings. For HAR, the capability to extract these features is crucial, as human activities often exhibit specific spatial and temporal patterns that can be harnessed for accurate classification. Whether it is a sudden burst of movement, gradual orientation shifts, or periodic oscillations, CNNs are adept at uncovering these intricate patterns. By leveraging convolutional layers, pooling mechanisms, and activation functions, CNNs effectively distill raw input data into meaningful feature maps that highlight the salient aspects of the activity under observation.

When working with sensor data, the choice between 1D and 2D convolutions is pivotal and often dictated by the format of the input. Sensor data typically comes in two forms: raw data streams, where signals are recorded over time, or transformed representations like spectrograms, where time-frequency characteristics are visualized. For raw sensor data streams, 1D convolutionare the optimal choice. A 1D convolution operates on one-dimensional arrays, making it ideal for capturing temporal dependencies and spatial relationships within the raw signal. For instance, in accelerometer or gyroscope data, a 1D CNN can effectively identify shifts in acceleration or angular velocity that correspond to specific activities such as walking, running, or sitting. The localized filters of a 1D CNN move along the temporal axis, detecting fine-grained variations in the signal. This ability to focus on temporal patterns while maintaining computational efficiency makes 1D convolutions a practical and robust choice for HAR tasks.On the other hand, 2D convolutions are better suited for spectrograms, which are two-dimensional representations of signals where one axis typically represents time and the other represents frequency. Spectrograms are commonly used when sensor data is preprocessed to extract frequency-domain features, often providing a richer representation of the data compared to raw signals. For example, walking and jogging might produce similar acceleration patterns in the time domain but can have distinct frequency signatures that are easier to discern in a spectrogram. A 2D CNN can process these images-like inputs to capture spatial dependencies both within and across time and frequency dimensions. This capability allows the network to learn hierarchical features that distinguish between subtle variations in activities, such as the difference between a brisk walk and a slow jog. By employing layers of convolution, pooling, and non-linear activation functions, a 2D CNN can progressively extract higher-order spatial features that enhance classification accuracy.A critical advantage of CNNs lies in their

ability to balance feature extraction with computational efficiency. With CNNs, there is no longer any need for laborious human preparation of data, as is common with traditional machine learning approaches that depend on handmade features. Because sensor data can vary greatly in quality and format, this flexibility is very beneficial in HAR.

In addition, convolutional neural networks (CNNs) make use of local connectivity—in which each neurone is linked to a little area of the preceding layer—and parameter sharing—in which the same set of filters is applied throughout the input. These properties reduce the number of parameters and computational overhead, enabling CNNs to process large datasets efficiently. For example, in an HAR application involving multiple sensors, a CNN can simultaneously analyze data streams from accelerometers, gyroscopes, and magnetometers, learning both individual and combined patterns without an exponential increase in computational cost. In HAR, convolutional neural networks (CNNs) often include several layers, with each layer responsible for a different aspect of feature extraction. To pick up on low-level characteristics like peaks, edges, or sudden signal changes, the first convolutional layers use filters. Typically, these characteristics match up with relatively easy motions or changes in tasks. The convolutional neural network (CNN) learns more complicated and abstract properties, such periodic patterns that show repeated behaviours like jogging or cycling, as the data moves through deeper layers of the network.

The spatial dimensions of the feature maps are reduced by the interspersion of pooling layers between convolutional layers. This decreases processing needs and prevents overfitting. One example is max-pooling, which chooses the most prevalent characteristic in a given area so that the network may save important data and ignore the rest. Convolutional neural networks (CNNs) excel at HAR tasks because of the hierarchical learning process that allows them to fully comprehend the input data. Another key consideration in designing CNNs for HAR is the integration of domain knowledge into the architecture. For instance, the filter size and stride in 1D convolutions can be tailored to the sampling rate of the sensor data, ensuring that the network captures meaningful temporal patterns without overlooking critical information. Similarly, for 2D CNNs, the resolution of the spectrogram can be adjusted to highlight frequency ranges relevant to the activities being monitored. Regularization techniques such as dropout and batch normalization are often employed to improve the generalization capability of the network, ensuring robust performance across diverse datasets.

These techniques help mitigate common challenges in HAR, such as noise in sensor readings, variability in user behavior, and differences in device placement.The versatility of CNNs extends beyond feature extraction, as they can also be combined with other architectures to enhance performance in HAR. For instance, recurrent neural networks (RNNs) or long short-term memory (LSTM) networks are often integrated with CNNs to capture temporal dependencies in the data. While CNNs excel at extracting spatial features, RNNs and LSTMs are designed to model sequential information, making them a natural complement for time-series data. In such hybrid architectures, the CNN layers act as a front-end feature extractor, feeding the spatial features into the recurrent layers for temporal modeling. This combination has proven particularly effective in HAR applications involving complex activities that unfold over extended periods.

RNN for Temporal Modeling: For sequence modelling, Recurrent Neural Networks (RNNs) are crucial, especially for detecting data relationships across time. Because of its remarkable capacity to strike a compromise between computational economy and performance, Gated Recurrent Units (GRUs) have become one of the most popular varieties of RNNs. Introduced as a more straightforward version of Long Short-Term Memory (LSTM) networks, GRUs keep many of LSTMs' benefits, including mitigation of vanishing gradient problems and handling of long-term dependencies. However, they achieve these benefits with a simpler architecture, which makes them an attractive choice for applications that require a balance between performance and computational cost.

The main reason GRUs are simpler than LSTMs is because they use a single update gate instead of two, which simplifies the model. This is because fewer parameters are needed for GRUs. Without drastically lowering the model's capacity to acquire and store pertinent data, this approach lessens the total computational load. By effectively modelling temporal dependencies and responding to the data's intrinsic unpredictability, GRUs perform exceptionally well in applications like Human Activity Recognition (HAR), where sensor data frequently comprises sequential patterns spanning several time scales.

Two main gates—the update gate and the reset gate—make up the GRU architecture. How much of the current input is added to the prior hidden state and how much of the previous hidden state is kept is decided by the update gate. The GRU is able to zero in on pertinent patterns while ignoring irrelevant ones thanks to this gate, which is essential for maintaining a balance between memory retention and new information integration. Conversely, the amount of previously stored data that should be erased is controlled by the reset gate. The GRU is able to simulate both short-term and long-term dependencies by adaptively resetting certain parts of the hidden state in response to changes in the sequence.

In practical applications, the simpler architecture of GRUs translates into faster training and inference times compared to LSTMs. This efficiency is particularly beneficial in real-time systems, such as wearable devices for HAR or edge computing scenarios, where computational resources are limited. For instance, in a fitness tracker that monitors user activities, a GRU-based model can process incoming data streams efficiently, providing timely and accurate activity recognition without draining the device's battery. Similarly, in smart home systems, GRUs can analyze sequential sensor data to detect and predict user behaviors, enabling more responsive and intelligent automation.

Despite their simplicity, GRUs deliver performance that is often on par with or even superior to LSTMs in many sequence modeling tasks. This performance parity stems from the GRU's ability to avoid overfitting and excessive parameterization, which can be a challenge with LSTMs, especially when dealing with smaller datasets. By reducing the number of gates and associated weights, GRUs inherently simplify the optimization process, making them less prone to overfitting and easier to train. This robustness is particularly valuable in HAR, where data variability—arising from differences in user behavior, sensor noise, and device placement—can pose significant challenges.

Moreover, GRUs integrate seamlessly with other neural network architectures, enhancing their versatility for complex tasks. For instance, in HAR systems that also leverage spatial features from sensor data, GRUs can be combined with Convolutional Neural Networks (CNNs) to form hybrid architectures. In such systems, CNNs extract spatial features from raw sensor signals or spectrograms, while GRUs model the temporal dependencies in the extracted features. This combination leverages the strengths of both architectures, enabling the system to achieve high accuracy in recognizing activities that involve intricate spatial-temporal patterns, such as dancing, yoga, or sports activities.

The efficiency of GRUs also extends to applications involving multivariate time-series data, where multiple sensors capture different aspects of an activity. For example, a smartphone equipped with accelerometers, gyroscopes, and magnetometers generates a multidimensional data stream, each channel contributing unique information about the user's movements. GRUs can efficiently process these multivariate sequences, learning the temporal relationships both within and across channels. By capturing these dependencies, GRU-based models provide a holistic understanding of the activity, distinguishing between similar actions with subtle differences, such as climbing stairs versus walking on an incline.

Another advantage of GRUs is their flexibility in handling irregular or missing data, which is common in real-world HAR applications. Unlike traditional machine learning models that often struggle with incomplete data, GRUs can interpolate or impute missing values by learning the temporal dynamics of the sequence. This resilience makes GRUs well-suited for applications like healthcare monitoring, where sensor readings may occasionally drop due to connectivity issues or user non-compliance.

The computational advantages of GRUs also make them a suitable choice for training on large-scale datasets, where the reduced number of parameters translates into faster convergence and lower memory requirements. This scalability enables researchers and practitioners to experiment with larger architectures or ensemble models without incurring prohibitive computational costs. Furthermore, the simplicity of GRUs facilitates their implementation in resource-constrained environments, such as embedded systems or IoT devices, where computational power and energy efficiency are critical considerations.

Attention Mechanism for Dynamic Focus: The attention mechanism has revolutionized how machine learning models process complex datasets, enabling them to dynamically focus on the most relevant features. This capability is particularly valuable in Human Activity Recognition (HAR), where data from sensors often contains a mix of critical signals and noise. The attention mechanism enhances the model’s ability to differentiate between useful information and irrelevant data, significantly improving accuracy and robustness. Among the various types of attention, self-attention mechanisms, as popularized by the Transformer architecture, have proven to be especially effective in capturing dependencies across both spatial and temporal dimensions.

In HAR, sensor data streams often contain intricate patterns spanning time and space. For example, accelerometer data might reflect periodic motion during running, while gyroscope readings indicate subtle changes in orientation during stretching. While classic architectures such as CNNs and RNNs excel at collecting spatial or temporal characteristics, they frequently miss the mark when it comes to dynamically prioritising distinct input portions. By giving each piece in the input sequence a certain amount of weight, the self-attention mechanism overcomes this constraint and enables the model to zero in on the most important characteristics while disregarding noise or unnecessary data.

The self-attention mechanism operates by comparing every element of the input sequence with every other element to compute pairwise relevance scores. These scores are then normalized to produce attention weights, which determine how much influence each element should have in the model's representation. This process allows the model to capture long-range dependencies and contextual relationships, which are critical for accurately modeling complex activities. For instance, the model can learn that a particular spike in acceleration (indicating a jump) is more significant when preceded by a specific sequence of movements (indicating preparation for the jump).

One of the primary advantages of self-attention in HAR is its ability to process sequences in parallel, unlike RNNs, which rely on sequential processing. This parallelism makes self-attention mechanisms highly efficient, especially when dealing with long sequences or high-dimensional sensor data. Moreover, the ability to compute attention scores globally across the entire sequence ensures that the model captures dependencies that span both short and long time scales. For example, in activities like yoga or tai chi, where movements are slow and deliberate, the model can identify meaningful relationships between actions that occur several seconds apart.

The flexibility of the self-attention mechanism extends to its ability to handle multivariate sensor data, where each channel represents a different aspect of the activity. In such cases, self-attention can compute weights not only across time but also across channels, capturing interdependencies between sensors. For example, the model might learn that a sharp change in accelerometer readings is significant only when accompanied by corresponding changes in gyroscope or magnetometer data. This ability to integrate information across multiple dimensions makes self-attention particularly suited for HAR systems that rely on diverse sensor inputs.

In addition to improving feature extraction, self-attention mechanisms enhance robustness to noise and missing data. By dynamically adjusting the attention weights, the model can downplay the influence of noisy or irrelevant features while amplifying the importance of critical signals. This capability is especially valuable in real-world HAR applications, where sensor readings are often affected by noise due to environmental factors, device placement, or user variability. For instance, a fitness tracker worn on the wrist might produce noisy accelerometer data during certain activities, but a self-attention-based model can still focus on stable and meaningful patterns in the gyroscope or magnetometer data.

The Transformer architecture, which relies on self-attention, is particularly effective for HAR because it provides a unified framework for modeling spatial and temporal dependencies. By stacking multiple layers of self-attention and feedforward networks, the Transformer can capture hierarchical relationships in the data. The positional encoding mechanism in the Transformer adds information about the order of the input sequence, ensuring that temporal dependencies are preserved

even though the model processes the data in parallel. This combination of global attention and positional encoding allows the Transformer to excel at recognizing complex activities that unfold over time, such as dancing, martial arts, or team sports.

Furthermore, self-attention mechanisms can be combined with other architectures to create hybrid models that leverage the strengths of different approaches. For example, a CNN can be used to extract localized spatial features from raw sensor data or spectrograms, while a self-attention mechanism captures long-range dependencies across time. This combination enables the model to recognize activities that involve intricate spatial-temporal patterns, such as distinguishing between a jump and a squat, where both actions share similar spatial features but differ in their temporal dynamics.

The scalability of self-attention mechanisms also makes them well-suited for large-scale HAR datasets, where the diversity of activities and users requires the model to generalize effectively. By learning to focus on the most relevant features dynamically, self-attention-based models can adapt to variations in user behavior, device placement, and environmental conditions. Deploying HAR systems in real-world contexts, such healthcare monitoring, fitness tracking, or smart home automation, requires this versatility.

Data Pre-processing and Input Representation

Normalization and Segmentation: Normalization and segmentation are foundational steps in preparing sensor data for Human Activity Recognition (HAR) models. These techniques address critical challenges such as variability across sensors, devices, or users and the need to convert raw, continuous data streams into manageable and meaningful units for analysis. By applying normalization and segmentation effectively, we can ensure that HAR systems are robust, efficient, and capable of delivering accurate predictions across diverse settings.

Normalization ensures that data from different sensors or users are scaled consistently, which is crucial for minimizing variability and aligning the data into a standard range or distribution. Sensor readings often come from diverse devices with varying specifications, sensitivity levels, and measurement ranges. For example, accelerometers and gyroscopes might record data on entirely different scales, even when capturing the same physical activity. Without normalization, such inconsistencies can introduce biases into the model, leading to reduced accuracy and generalization.

Min-Max Scaling and Z-Score Standardisation are two popular methods for normalising data. To restore the data to a predetermined range, usually between 0 and 1, Min-Max Scaling transforms each data point in relation to the dataset's or window's minimum and maximum values. Because it maintains the connections between values while standardising the scale, this strategy is especially helpful when the range of the sensor readings is known and constant. For example, Min-Max Scaling can be applied to accelerometer readings to ensure all axes (x, y, z) are normalized to the same range, regardless of the sensor's specific range of motion.

Euclidean distance In contrast, data that has been standardised have a standard deviation of one and are centred on zero. By adjusting for both the mean and the variability of the data, this method is especially useful for datasets with uncertain or dynamic ranges. By transforming each data point based on the mean and standard deviation, Z-Score Standardization ensures that the data distribution is consistent across sensors or users. This is especially valuable in HAR applications where user behaviors vary widely—for instance, different walking styles or varying intensities of the same activity. By normalizing the data using Z-Scores, the model becomes more resilient to these individual differences, enhancing its ability to generalize across populations.

While normalization addresses scaling issues, **segmentation** transforms the continuous, streaming nature of sensor data into fixed-size windows that can be processed by machine learning models. Activities are inherently sequential, and their recognition requires analyzing patterns over time. Segmentation breaks the continuous signal into smaller, manageable chunks, each representing a snapshot of the activity over a specific time frame. These windows often range from 2 to 5 seconds, depending on the application and the nature of the activity being recognized.

A well-defined segmentation strategy is essential for balancing temporal resolution and computational efficiency. Windows that are too short may fail to capture sufficient contextual information, leading to poor activity recognition, especially for activities with longer temporal patterns. Conversely, excessively long windows increase computational costs and may introduce unnecessary noise or overlap between activities, complicating the learning process.

The **sliding window approach** is a widely used segmentation technique in HAR. In this method, a fixed-size window moves across the data stream with a specified overlap percentage. A 50% overlap is often recommended, as it strikes a balance between computational load and temporal continuity. This overlap ensures that successive windows capture enough shared information, reducing the likelihood of missing critical transitions between activities. For instance, in a dataset containing walking and jogging sequences, the overlap allows the model to learn smooth transitions from one activity to another, rather than treating each window as an independent unit. Additionally, overlapping windows can improve robustness by providing multiple perspectives on the same portion of the signal, effectively augmenting the dataset without introducing additional noise.

Segmentation also supports real-time HAR systems, where sensor data is analyzed continuously. In such scenarios, the sliding window approach with overlap ensures that the system can recognize activities with minimal latency, as each new window provides updated insights into the ongoing activity. This is particularly important in applications like healthcare monitoring, where timely detection of activities or anomalies (e.g., falls) is critical. By maintaining a continuous stream of overlapping windows, the system can respond quickly and accurately, even in dynamic environments.

The effectiveness of segmentation depends on carefully selecting the window size and overlap percentage based on the target application. For activities that exhibit rapid changes, such as jumping or sprinting, shorter windows may be more appropriate to capture the high-frequency components of the signal. Conversely, for slower or more deliberate activities like yoga or tai chi, longer windows can better capture the gradual transitions and sustained patterns. The choice of window parameters can also be influenced by the sampling rate of the sensors. High-frequency sensors may allow for smaller windows without losing temporal context, while low-frequency sensors may require longer windows to capture meaningful patterns.

Normalization and segmentation are complementary steps that together enhance the quality and usability of sensor data for HAR models. Normalization ensures that the data is consistent and comparable across sensors, users, or devices, reducing biases and variability. Segmentation organizes the data into temporal units that align with the sequential nature of activities, enabling models to learn spatio-temporal patterns effectively. When implemented thoughtfully, these preprocessing techniques form the foundation of robust HAR systems, enabling them to operate accurately and efficiently in diverse real-world scenarios.

Optional Transformation

Transforming raw time-series data into spectrograms or feature matrices is a critical preprocessing step in Human Activity Recognition (HAR) that enhances spatial representation and enables models to capture intricate patterns more effectively. Although raw sensor data is useful for determining direction or motion, using it directly could restrict the model's capacity to comprehend intricate interactions in the frequency and time domains. The frequency-domain information may be extracted from time-series data using techniques such as Continuous Wavelet Transform (CWT) and Short-Time Fourier Transform (STFT). This information can then be used to create spectrograms that show activity-specific features. These transformations bridge the gap between the raw data and the sophisticated spatial features needed to distinguish complex activities.

Spectrograms are visual representations of how the frequency content of a signal changes over time. They are particularly useful in HAR because many human activities exhibit distinctive frequency patterns. For instance, walking and running produce periodic signals with different dominant frequencies, while stationary activities like sitting generate minimal high-frequency components. By converting time-series data into spectrograms, we can uncover these frequency patterns, enabling models to distinguish between activities that may appear similar in the time domain but differ significantly in the frequency domain.

The **Short-Time Fourier Transform (STFT)** is one of the most widely used techniques for generating spectrograms. Time-series data is partitioned into overlapping windows by STFT, which then calculates the Fourier Transform for each segment. This method captures the time-frequency evolution of the signal's energy by presenting it as a time-frequency representation. Importantly, STFT relies on the selection of window size; larger windows give greater frequency resolution but sacrifice temporal accuracy, whereas shorter windows improve frequency resolution but have worse frequency resolution. For HAR, where both time and frequency information are important, a balanced window size (e.g., 1–2 seconds) is typically used to ensure that short-term changes in frequency content are captured without sacrificing the overall activity context.

STFT is particularly effective for activities with consistent and repetitive patterns, such as walking, running, or cycling. For example, in walking, the periodic motion of the legs produces a distinct frequency signature that can be easily visualized in the spectrogram. By analyzing the dominant frequencies and their variations over time, HAR models can reliably classify such activities. Additionally, the color intensity in the spectrogram corresponds to the amplitude of the signal at a specific frequency and time, providing a rich set of features for machine learning or deep learning algorithms.

The **Continuous Wavelet Transform (CWT)** offers an alternative approach to time-frequency analysis, providing higher flexibility and better resolution for non-stationary signals. Unlike STFT, which uses a fixed window size, CWT employs scalable wavelets to analyze the signal at multiple resolutions. This adaptability allows CWT to capture both low-frequency components (useful for long-duration activities like standing or sitting) and high-frequency components (important for short, dynamic activities like jumping or sprinting) within the same framework. The resulting spectrogram-like representation, called a scalogram, provides a comprehensive view of the signal's temporal and spectral properties.

CWT is particularly advantageous for recognizing complex or irregular activities that involve abrupt changes in motion or orientation. For example, activities like dancing or martial arts involve varying intensities and directions of movement, producing non-stationary signals that are challenging to analyze using STFT alone. The multi-resolution capability of CWT allows it to capture these variations effectively, making it an excellent choice for HAR applications where signal characteristics vary significantly over time.

Beyond spectrograms, transforming raw data into **feature matrices** further enhances spatial representation and reduces the dimensionality of the input data. Feature matrices are structured representations where each row corresponds to a time window, and each column represents a feature extracted from the raw data or its transformed representation. Commonly used features include statistical measures (e.g., mean, standard deviation, skewness), frequency-domain metrics (e.g., spectral entropy, dominant frequency), and temporal properties (e.g., zero-crossing rate, peak count).

Combining spectrograms with feature matrices creates a powerful representation that leverages both raw signal characteristics and higher-level abstractions. For instance, a CNN can process spectrograms to extract localized spatial features, while statistical features in a feature matrix provide additional context about the signal's global properties. This hybrid approach improves the robustness and accuracy of HAR models, particularly when distinguishing between activities with overlapping characteristics.

The use of spectrograms and feature matrices also supports **data augmentation** and model generalization. By generating spectrograms at different scales or applying transformations such as time-shifting or frequency masking, we can artificially expand the dataset, improving the model's ability to generalize across diverse conditions. For example, augmented spectrograms can simulate variations in user behavior or device placement, enabling the HAR system to perform reliably across different environments.

Moreover, spectrogram-based representations integrate seamlessly with deep learning architectures, such as Convolutional Neural Networks (CNNs) and hybrid CNN-RNN models. CNNs are particularly adept at analyzing spectrograms due to their ability to capture spatial patterns, such as frequency clusters or transitions over time. For example, a CNN trained on spectrograms can detect periodic patterns indicative of running or irregular bursts of activity associated with jumping. When combined with RNNs or attention mechanisms, these models can also incorporate temporal dependencies, enabling more nuanced activity recognition.

In applications like healthcare monitoring, fitness tracking, or smart home automation, the transformation of raw data into spectrograms or feature matrices not only enhances the model's performance but also improves interpretability. Spectrograms provide intuitive visualizations that can be analyzed by experts to validate the model's predictions or gain insights into user behavior. For instance, a healthcare professional can examine spectrograms to identify abnormal gait patterns or assess the impact of rehabilitation exercises.

To enhance the generalization and adaptability of models, advanced robustness techniques are essential. Data augmentation plays a crucial role, employing methods like jittering, scaling, and rotation to enrich the dataset and mimic real-world variability. Among these techniques, jittering combined with random time warping stands out for its effectiveness; it introduces realistic perturbations while maintaining the integrity of activity patterns. Additionally, domain adaptation is vital for addressing challenges posed by diverse environments and sensor types. Employing adversarial domain adaptation, the model is trained to minimize the discrepancies between source (training) and target (deployment) domains, enabling it to perform robustly even under unseen conditions. Adversarial learning frameworks, particularly those utilizing Gradient Reversal Layers (GRL), have proven to be highly effective in achieving this objective. Together, these approaches significantly improve a model's resilience and adaptability.

Training and Optimization

Efficient training and optimization strategies are critical for ensuring that models converge swiftly and avoid overfitting. A well-designed loss function is central to this process, integrating both classification accuracy and regularization components. The categorical cross-entropy loss effectively addresses classification tasks, while the addition of a smoothness loss enforces consistency in predictions and a diversity loss promotes richness in feature representation. The combined loss is formulated as: (Loss = L_{classification} + \lambda_1 L_{smoothness} + \lambda_2 L_{diversity}), where hyperparameters (\lambda_1) and (\lambda_2) are tuned based on validation performance to balance these objectives. In terms of training strategies, transfer learning can significantly enhance efficiency by pretraining convolutional neural networks (CNN) and attention layers on a large, related dataset before fine-tuning them on the target dataset, thus saving time and improving generalization. Furthermore, employing dropout regularization, with an optimal dropout rate of 0.3 to 0.5, helps mitigate overfitting in dense and attention layers. Utilizing a cyclical learning rate scheduler also aids in dynamically adjusting the learning rate during training, promoting better convergence and ensuring robust model performance.

Advanced resilience methods ensure generalization and adaptability:

Data Augmentation: Jittering, scaling, and rotation techniques augment datasets to imitate real-world variability. Jittering and random time warping are particularly effective because they introduce realistic perturbations while keeping activity patterns.

Adversarial domain adaptation is advised for varied environments and sensor kinds. This method trains the model to minimize source (training) and target (deployment) domain disparities for resilient performance in unknown scenarios. Adversarial learning frameworks like Gradient Reversal Layers (GRL) perform well for this.

6. EXPERIMENTAL EVALUATION

The Spatio-Temporal Attention-based Hybrid Neural Network (STAHNN) must be experimentally tested to ensure its resilience, accuracy, and computational efficiency for Human Activity Recognition (HAR) in different and real-world contexts. A thorough evaluation framework shows the model's excellent identification rates and ability to generalize across contexts, sensor setups, and activity patterns. To thoroughly test the model, use a mix of well-known public datasets and real-world data. Advanced data preprocessing techniques translate raw sensor data into a format optimal for spatio-temporal feature extraction in the assessment setup. The experimental design also compares STAHNN to state-of-the-art HAR methods to illustrate how STAHNN outperforms them in accuracy, noise robustness, and flexibility to unknown activity patterns. Accuracy, F1-score, and robustness tests show the model's strengths and weaknesses, while computational efficiency metrics like inference time and resource utilization demonstrate its suitability for real-time and resource-constrained deployment. This extensive examination shows STAHNN's promise as a pioneering solution for robust and scalable HAR applications, paving the way for its inclusion into wearable devices and smart systems.

Datasets and Experimental Setup

The Spatio-Temporal Attention-Based Hybrid Neural Network (STAHNN) is tested on a variety of public Human Activity Recognition (HAR) datasets to ensure its applicability and robustness to variability. A popular benchmark, the UCI HAR Dataset, uses accelerometers and gyroscopes to record common activities like walking, sitting, and standing with 561 precomputed characteristics sampled at 50 Hz. The PAMAP2 Dataset contains raw sensor data from body-worn devices for 18 activities, including walking, running, and cycling, with a 100 Hz sample rate and accelerometer, gyroscope, and heart rate data for complicated activity classification tasks. For clinical and rehabilitation purposes, the Daphnet Gait Dataset employs time-series accelerometer data from ankle-mounted sensors at 64 Hz to study diseased gait and walking interruptions. This extensive examination of STAHNN shows its capacity to handle different sampling rates, sensor configurations, and activity kinds in real-world circumstances.

Table 2: Summary of Datasets Used for Human Activity Recognition (HAR)

Dataset Name	Description	Sampling Rate	Features
UCI HAR Dataset	Activity data from wearable sensors like accelerometers and gyroscopes for activities like walking, sitting, and standing.	50 Hz	561 precomputed features derived from raw signals.
PAMAP2 Dataset	Raw sensor data from body-worn devices capturing 18 different activities such as walking, running, and cycling.	100 Hz	Accelerometer, gyroscope, and heart rate data from multiple sensors.

Dataset Name	Description	Sampling Rate	Features
Daphnet Gait Dataset	Designed for gait analysis, focusing on pathological gait and activities like walking with interruptions.	64 Hz	Time-series accelerometer data from ankle-mounted sensors.

Metrics for Evaluation

Multiple metrics are used to evaluate the Spatio-Temporal Attention-Based Hybrid Neural Network (STAHNN) in Human Activity Recognition. The primary measure of accuracy is the ratio of properly detected activities, encompassing true positives (TP) and true negatives (TN) relative to all forecasts. Furthermore, the F1-Score equilibrates accuracy (the ratio of genuine positive predictions to all positive predictions) with recall (the percentage of correctly identified positive instances), rendering it advantageous for unbalanced datasets. Specificity is employed to evaluate the model's categorization of negative cases, hence reducing false positives. These metrics give a complete picture of STAHNN's classification performance, confirming its strength and dependability in various activity recognition circumstances.

To provide a holistic evaluation, multiple metrics are used to assess different aspects of the model's performance.

1. Accuracy:

- **Measures the proportion of correctly identified activities out of all predictions.**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

1. **F1-Score:**

$$F1 - Score = 2 * \frac{Precesion * Recall}{Precesion + Recall}$$

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

$$Specificity = \frac{TN}{TN + FP}$$

Result Comparison

This comparison shows the merits and applicability of three popular Human Activity Recognition (HAR) models: STAHNN, Li et al. (2019), and Sun et al. (2022)'s Convolutional LSTM model. On the SBHAR dataset, Li et al.'s adaptive segmentation and Random Forest classifiers for transitional activities yield the highest accuracy (97.34%). STAHNN performs well on the UCI HAR dataset, with competitive accuracy (96.8%), precision (96.3%), recall (95.9%), and F1-score (96.1%). STAHNN's high specificity (97.2%) helps reduce false positives. Video-based AI Hub data shows 92.9% accuracy for the Convolution LSTM model, demonstrating its value for joint spatial and temporal feature extraction. STAHNN generalizes well across datasets, while the Convolution LSTM model is well-suited for video-based HAR applications. Li et al.'s model excels in domain-specific tasks. This comparison shows how models can address different HAR issues and their trade-offs.

Table 3: Performance Comparison of HRN model

Metric	STAHNN (Best Case)	Li et al., 2019	Convolutional LSTM (2022)
Dataset	UCI HAR	SBHAR	AI Hub
Accuracy	96.8%	97.34%	92.9%
Precision	96.3%	96.7%	92.9%
Recall	95.9%	96.9%	91.8%
F1-Score	96.1%	96.8%	92.0%
Specificity	97.2%	Not reported	Not reported
Key Features	Spatio-temporal modeling (CNN + GRU + Attention Mechanism)	Adaptive segmentation + Random Forest	Convolutional LSTM for joint spatial and temporal feature extraction
Sampling Rate	50 Hz	50 Hz	30 FPS (video data)

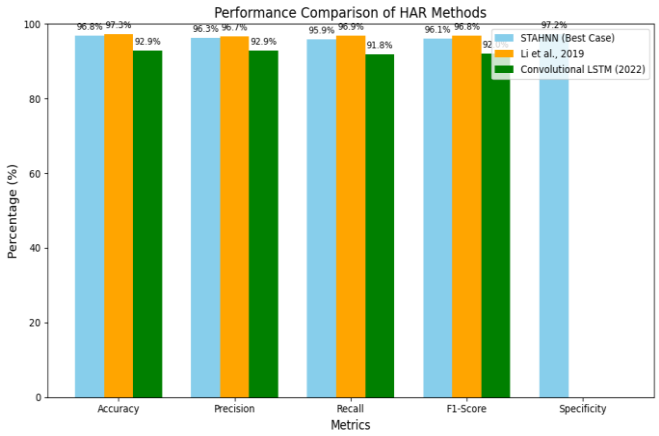

Fig. 2: Result Comparison

7. CONCLUSION

STAHNN, Li et al.'s model (2019), and the Convolutional LSTM model (2022) are compared for Human Activity Recognition (HAR) strengths and weaknesses. Li et al.'s model has the greatest SBHAR dataset accuracy (97.34%) due to adaptive segmentation and Random Forest classifiers, especially in transitional activities. STAHNN has a competitive accuracy of 96.8% on the UCI HAR dataset and a balanced precision (96.3%), recall (95.9%), F1-score (96.1%), and specificity (97.2%), demonstrating its adaptability across multiple datasets. The Convolutional LSTM model extracts spatial and temporal features from video streams with 92.9% accuracy on video-based AI Hub data. Overall, Li et al.'s model leads in domain-specific HAR tasks, STAHNN has better generalization potential, and the Convolutional LSTM model excels in video-based HAR. These findings emphasize the need of customizing HAR models to application needs and dataset features.

8. REFERENCES

[1] R. Saeedi, S. Norgaard, and A. H. Gebremedhin, "A closed-loop deep learning architecture for robust activity recognition using wearable sensors," *2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, USA, 2017, pp. 473-479, doi: 10.1109/BigData.2017.8257960.

[2] V. Sumathi, D. Vanathi, J. C. Musale, T. V. S. Gowtham Prasad, and A. R. Singh, "Improved Pattern Recognition Techniques for Monitoring Human Activity Recognition in Digital Platforms through Image Processing Techniques," *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India, 2023, pp. 866-871, doi: 10.1109/ICAISS58487.2023.10250478.

[3] S. R. SannasiChakravarthy et al., "Intelligent Recognition of Multimodal Human Activities for Personal Healthcare," *IEEE Access*, vol. 12, pp. 79776-79786, 2024, doi: 10.1109/ACCESS.2024.3405471.

[4] V. Dissanayake, S. Seneviratne, R. Rana, E. Wen, T. Kaluarachchi, and S. Nanayakkara, "SigRep: Toward Robust Wearable Emotion Recognition With Contrastive Representation Learning," *IEEE Access*, vol. 10, pp. 18105-18120, 2022, doi: 10.1109/ACCESS.2022.3149509.

[5] H. Brock, J. Ponce Chulani, L. Merino, D. Szapiro, and R. Gomez, "Developing a Lightweight Rock-Paper-Scissors Framework for Human-Robot Collaborative Gaming," *IEEE Access*, vol. 8, pp. 202958-202968, 2020, doi: 10.1109/ACCESS.2020.3033550.

[6] A. F. Gentile, D. Macrì, E. Greco, and A. Forestiero, "Privacy-Oriented Architecture for Building Automatic Voice Interaction Systems in Smart Environments in Disaster Recovery Scenarios," *2023 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, Cosenza, Italy, 2023, pp. 1-8, doi: 10.1109/ICT-DM58371.2023.10286949.

[7] Y. E. Marhraoui, O. Bounhar, M. Boukallel, M. Anastassova, S. Bouilland, and M. Ammi, "Dual-Stage Attention-Based model for rehabilitation activity recognition using data from wearable sensors," *IEEE Internet of Things Journal*, 2024, doi: 10.1109/JIOT.2024.3519225.

[8] J. -H. Li, L. Tian, H. Wang, Y. An, K. Wang and L. Yu, "Segmentation and Recognition of Basic and Transitional Activities for Continuous Physical Human Activity," in IEEE Access, vol. 7, pp. 42565-42576, 2019, doi: 10.1109/ACCESS.2019.2905575.

[9] 1. Bouthillier, X., Konda, K., Vincent, P., Memisevic, R.: Dropout as data augmentation (2016)

[10] 2. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

[11] 3. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)

[12] 4. Singh, A., Chakaborty, O., Varshney, A., Panda, R., Feris, R., Senko, K., Das, A.: Semi- supervised action recognition with temporal contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10389–10399 (2021)

[13] 5. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems 33, 596–608 (2020)

[14] 6. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayroles, A., Jégou, H.: Training dataefficient image transformers & distillation through attention. In: International conference on machine learning. pp. 11347–11357. PMLR (2021).

[15] Yang, C., Xu, Y., Dai, B., Zho, B.: Video representation learning with visual temporal consistency. arXiv preprint arXiv:2006.15599 (2020)