


# HSTCN-NuSVC: A Homogeneous Stacked Deep Ensemble Learner for Classifying Human Actions Using Smartphones

Sarmela Raja Sekaran <sup>1</sup>, Ying Han Pang <sup>1\*</sup>, Ooi Shih Yin <sup>1</sup>, Lim Zheng You <sup>1</sup>

<sup>1</sup> Faculty of Information Science and Technology, Multimedia University, Malacca, Malaysia.

## Abstract

Smartphone-based human activity recognition (HAR) is an important research area due to its wide-ranging applications in health, security, gaming, etc. Existing HAR models face challenges such as tedious manual feature extraction/selection techniques, limited model generalisation, high computational cost, and inability to retain longer-term dependencies. This work aims to overcome the issues by proposing a lightweight, homogenous stacked deep ensemble model, termed Homogenous Stacking Temporal Convolutional Network with Nu-Support Vector Classifier (HSTCN-NuSVC), for activity classification. In this model, multiple enhanced TCN networks with diverse architectures are organised parallelly to capture hierarchical spatial-temporal patterns from raw inertial signals. Each base model (i.e., TCN) incorporates dilations and residual connections to preserve longer effective histories, allowing the model to retain longer-term dependencies. Additionally, dilations can diminish the number of trainable parameters, reducing the model complexity and computational cost. The base models' predictions are concatenated and fed into a meta-learner (i.e., Nu-SVC) for final classification. The proposed HSTCN-NuSVC is evaluated using a publicly available database, i.e., UCI HAR, and a subject-independent protocol is implemented. The empirical results demonstrate that HSTCN-NuSVC achieves 97.25% accuracy with only 0.51 million parameters. The results exhibit the model's effectiveness in enhancing generalisation across individuals with better accuracy and computational efficiency.

## Keywords:

Deep Ensemble Learning;  
Smartphone-Based Human Activity Recognition;  
Stacking Ensemble;  
Lightweight Model;  
Hierarchical Deep Features.

## Article History:

Received:	13	June	2024
Revised:	14	January	2025
Accepted:	18	January	2025
Published:	01	February	2025

## 1- Introduction

### 1-1- Background

In today's rapidly changing world, researchers and scientists are constantly conducting research and introducing new technological innovations. These advancements have profoundly influenced daily life, making it safer, easier, and more efficient. One of the most sought-after research areas is Human Activity Recognition (HAR), which plays a pivotal role in improving the quality of life. HAR involves the use of computers or machines to automatically identify human motions and actions based on input data collected from various devices, such as cameras, wearable sensors, and smartphones. Thanks to its autonomous motion recognition capabilities, HAR has found applications in areas like patient rehabilitation, ambient-assisted living, geriatric care, personal physical activity monitoring, and security systems [1, 2]. For instance, employing the HAR system in a hospital can reduce the manual labour for continuous patient monitoring.

Human Activity Recognition (HAR) systems are categorised into vision-based and non-vision-based approaches, depending on the type of data used for classifier training. Vision-based HAR models rely on two-dimensional images or three-dimensional videos, while non-vision-based models use one-dimensional sensor signals. Due to their broad

\* **CONTACT:** [yhpang@mmu.edu.my](mailto:yhpang@mmu.edu.my)

**DOI:** <http://dx.doi.org/10.28991/ESJ-2025-09-01-026>

© 2025 by the authors. Licensee ESJ, Italy. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<https://creativecommons.org/licenses/by/4.0/>).

applicability in areas such as video surveillance, security, virtual reality technology, and human-robotic interaction, vision-based HAR is widely adopted [3–5]. It has gained significant attention within the computer vision community. However, vision-based Human Activity Recognition (HAR) encounters several challenges. For instance, this approach demands substantial computational resources during both data pre-processing and the training phase, owing to the complexity of the input samples. Moreover, the performance of these HAR models is highly sensitive to background interference (such as non-human movements), varying lighting conditions, and occlusions [6]. Moreover, the volunteers are at risk of privacy violations during data collection since user identification attributes are collected. These make non-vision-based HAR solutions more favourable. Unlike vision-based HAR, non-vision-based HAR models are relatively lightweight, resistant to background noises, inconsistent lighting, and occlusions, and preserve user privacy since the collected data are one-dimensional motion signals [6, 7]. Additionally, the data collection process in non-vision-based HAR is non-intrusive and relatively more comfortable for the volunteers than in vision-based HAR. This is because the volunteers only need to carry or wear a smartphone or smartwatch while performing activities without any identity disclosure.

### ***1-2-Related Work***

Designing robust HAR systems that accurately identify human activities and behaviours is pivotal, as these systems can be integrated into real-life applications such as physical activity tracking, fall detection, sedentary detection, etc. As mentioned, handcrafted feature-based and deep learning approaches are extensively used in the smartphone-based HAR domain. In earlier times, HAR solutions were based on handcrafted feature-based methods due to the scarcity of computational resources. The data analysts studied the nature and characteristics of human motion data and developed feature engineering techniques to extract meaningful information that differentiates between activities. Generally, the computed features will undergo ranking and selection processes to eliminate redundancy. Then, an appropriate machine learning classifier will be chosen for activity classification. Furthermore, the machine learning classifier performance can be enhanced through parameter optimisation. Handcrafted feature-based models achieved adequate performances, albeit lacking in generalising capabilities [8-10]. For example, Anguita et al. [11] gathered inertial signals from thirty volunteers and performed manual statistical feature extraction to compute 563 features. Then, the author selected a multiclass Support Vector Machine (SVM) to classify the extracted statistical features. Similarly, Batool et al. [7] calculated statistical features and ran them through the Particle Swarm Optimisation (PSO) algorithm to further select desired features before SVM classification. Ansari et al. [8] introduced a feature selection technique inspired by the wrapper and filter methods to enable sequential floating forward search before passing the extracted feature into the SVM model. This enables the model to discard redundant and irrelevant features.

Another popular handcrafted technique is the K-Nearest Neighbour (KNN) model, which classifies human activities. Tharwat et al. [6] trained the KNN model with normalised input signals and identified a suitable K value for the classifier using the PSO algorithm. The authors claimed that the PSO algorithm reduced the error rates compared to the other optimisation algorithms, such as the Genetic and Ant Bee Colony Optimisation algorithms. Mohsen et al. [12] also implemented the KNN algorithm to identify activity classes. The authors tested the model with varying K values and deduced that increasing the K value enhances the classification performance. Besides the KNN model, Kee et al. [13] developed a Random Forest (RF) classifier and evaluated the model using the subject-independent protocol where the training and testing sets have different user samples. The authors found that their model performed better than existing standard machine learning classifiers by achieving 83.3% accuracy.

Researchers have focused more on deep learning solutions, including convolutional models [14, 15], recurrent models [16, 17], and Temporal Convolutional Networks (TCN) [18, 19], to solve the HAR tasks because these methods can autonomously capture desirable features from the motion signals without human involvement in recent years. Most deep learning architectures can be directly trained on raw input signals since these models perform autonomous deep feature extraction. For example, Han et al. [15] developed a heterogenous CNN using grouped convolutions with varying kernel sizes to better extract global contextual information representing each activity class. In the work of Liu et al. [20] and Khan et al. [21], a CNN model consisting of parallelly organised multiple convolutional streams was introduced, each with an attention mechanism called the Squeeze and Excite module for removing redundant details. The empirical results demonstrated the efficacy of this model in improving the model's performance. Likewise, Hamad et al. [14] designed three-layered one-dimensional dilated causal convolutional layers followed by a multi-head self-attention module. The convolutional layers capture temporal information, and the self-attention module retains salient information and removes irrelevant features. Sharen et al. [22] developed a convolutional architecture (WISNet) comprising three blocks, including Convolved Normalised Pooled (CNP<sub>M</sub>), Identity and Basic (IDB<sub>N</sub>) and Channel and Spatial Attention (CAS<sub>b</sub>) blocks. CNP<sub>M</sub> captures relevant features from input signals, IDB<sub>N</sub> is responsible for extracting residual progressive features that capture complex sequential data differences, and CAS<sub>b</sub> eliminates irrelevant information based on relative weights. The authors claim that WISNet outperformed other models by achieving 96.41% on WISDM, 95.66% on UCI HAR, and 94.01% on KU-HAR.

As for recurrent models, the most commonly used models are LSTM [17, 22, 23] and Gated Recurrent Unit (GRU) [18, 24, 25] networks. For instance, Sharen et al. [22] proposed a stacked LSTM architecture that outperformed the other machine learning models by approximately 6%. Dubey et al. [26] built a hybrid architecture using multiple convolutional layers and stacked LSTM layers to extract spatiotemporal features from motion signals. This model attained a 93.5 F1 score on UCI HAR with 0.46 million parameters. Bidirectional LSTM is another variant of the LSTM model that could extract information in both directions during model training. In other words, this model is able to process previous and subsequent information simultaneously to make predictions. Hence, Hernandez et al. [16] adopted this architecture and built a three-layer bidirectional LSTM layer to perform activity recognition. Although the bidirectional LSTM model had 92.67% accuracy on the UCI HAR database, this model struggled to differentiate between two static activities (i.e., sitting and standing classes). Besides LSTM networks, Pan et al. [17] introduced a GRU-induced dual attention mechanism (i.e., channel and temporal attention modules), where these dual attention mechanisms reduce model biases and gather significant temporal information from the inertial signals. From the empirical results, it is noticed that implementing dual attention could improve the overall classification performance by approximately 3%. Furthermore, the authors claimed that their proposed model outperformed the existing methods, including SVM variants, KNN, Light GBM, and Random Forest (RF).

Since recurrent models are heavy in computations, Lai et al. [27] developed two new architectures, Dilated TCN and Encoder-Decoder TCN (ED-TCN), to perform human motion classification using video data. The authors reported that these models achieved better accuracy with relatively low computation. Owing to the benefits of TCN models, Nair et al. [18] adopted both models into the smartphone-based HAR domain to test the efficacy of the models. The authors found that TCN architectures outperformed the existing complex recurrent models due to their ability to capture longer temporal information from the input motion signals. Raja Sekaran et al. [19] introduced a new TCN variant inspired by Inception architecture called Multiscale Temporal Convolutional Network (MSTCN). MSTCN excels at performing multiscale feature extraction on the inertial signal, enabling the gathering of richer information. The authors incorporated L1, L2, and dropout regularisation into the model to increase the model's sparseness, reducing the overfitting effects. The experimental results demonstrated that MSTCN achieved 97.42% accuracy on UCI HAR with 3.75 million parameters.

Along with handcrafted and deep learning methods, scientists are increasingly promoting ensemble learning techniques when attempting smartphone-based HAR tasks. The preference is primarily due to the ability of ensemble learning techniques to leverage the strengths of various models, regardless of their type, contributing to superior classification performance. For example, Bhattacharya et al. [28] proposed Ensem-HAR, an ensemble of multiple deep learning architectures, using one-dimensional CNN and LSTM models [28]. The authors concatenated the predictions from the individual models and used them to train a blender for the final prediction. Similarly, Tan et al. [29] integrated CNN, GRU, and deep neural networks (DNN) to form an ensemble-learning algorithm (ELA) for motion feature extraction. The ELA model outperformed other compared models, with an accuracy of approximately 96%. Given the ensemble deep learning models' benefits and TCNs' exemplary performance in time series classification, we tackle the ensemble learning techniques, particularly stacking techniques, to reap the strengths of diverse TCN architectures in this study. Our objective is to introduce an efficient human activity recognition model that optimises the potential values of TCN diversity.

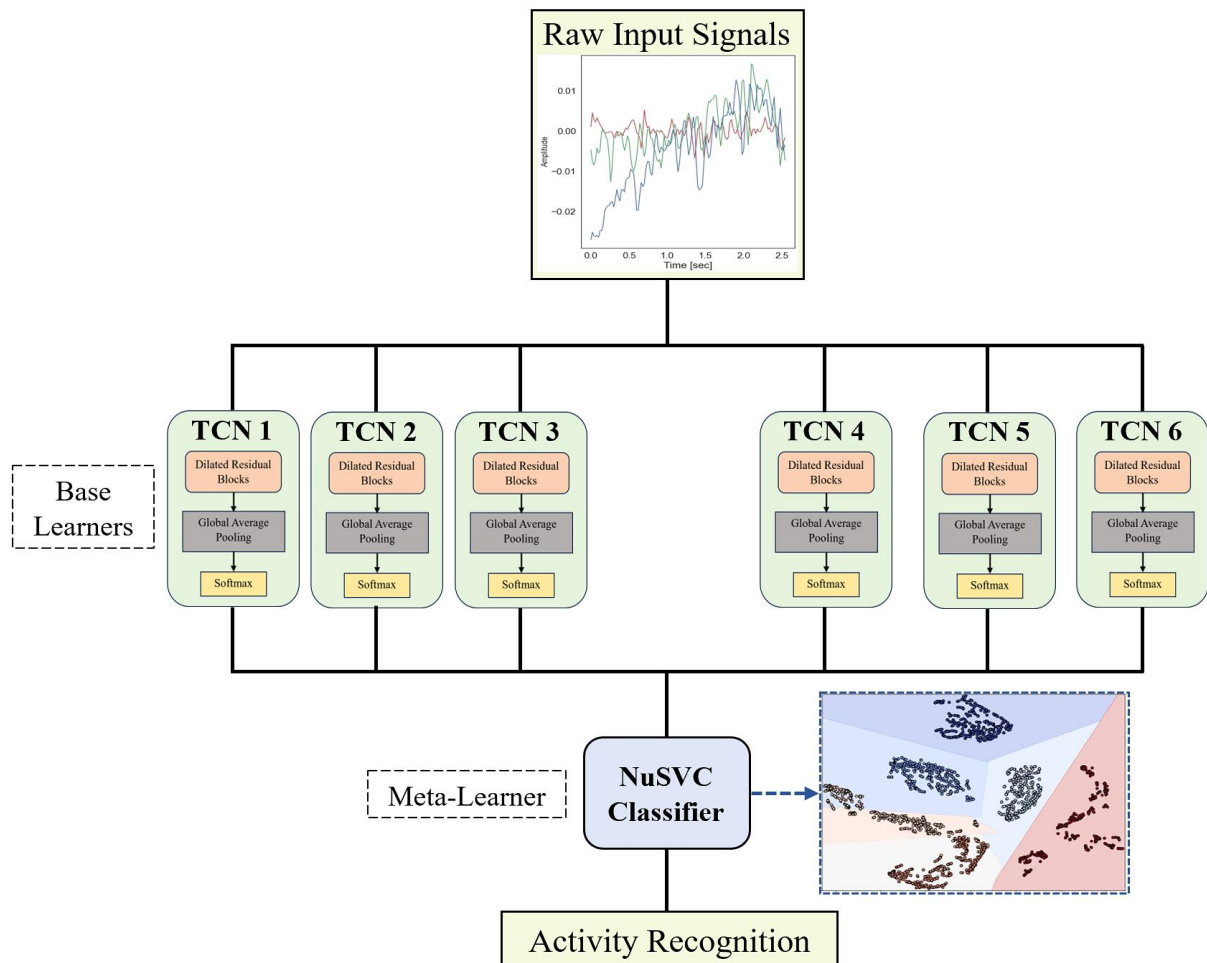
### *1-3-Motivations and Contributions*

As previously mentioned, though countless techniques have been developed, there are still several notable limitations in existing methods. The existing limitations are:

- Handcrafted feature-based methods require hardcore signal pre-processing and time-consuming manual feature engineering and selection techniques to enhance performance. No universally accepted standards for choosing optimal underlying patterns, considering that they differ between studies. Moreover, the manually crafted features rely on existing domain knowledge and may cause the classifier to overlook the implicit patterns of the inertial signals, negatively affecting the model's performance [14].
- Although deep learning models do not require labour-intensive feature extraction techniques, most of these models must be trained on large and diverse datasets to acquire optimal classification performance. For instance, a large and complex deep learning model trained on smaller training data is prone to overfitting, causing poorer classification accuracy to unseen data.
- Classical Convolutional Neural Networks (CNN) attain subpar performance in time series analysis as these models are inefficient at acquiring temporal features from the input motion signals [26].
- Though recurrent models like Long Short-Term Memory (LSTM) models are popular in time series classification tasks as they extract time-dependent information from the input samples, which is significant in gait/motion analysis, they are complex and require high processing time and computational resources for model training [30].
- Even though the TCN architecture, particularly Dilated TCN models, requires lower computation resources for model training, this model needs to be deeper for optimal recognition performance. This can be accomplished by incorporating larger, longer filters for deep temporal feature extraction. However, this may lead to overfitting.

Considering the above limitations, this research proposes a deep ensemble architecture, Homogeneous Stacked Temporal Convolutional Network with Nu-Support Vector Classifier (HSTCN-NuSVC). Stacking multiple TCN models with a meta-learner can be beneficial since it grants better model generalisation, leading to superior performance. To be specific, multiple TCN models with varying architectural parameters are trained, and the predictions from each model are concatenated and used to train the meta-model, NuSVC. Figure 1 provides an overview of the proposed HSTCN-NuSVC. Each TCN model in the stack, configured with different hyperparameters, captures a range of temporal characteristics at various levels of abstraction. This enables the extraction of complementary information from the inertial data, offering a more comprehensive analysis. Moreover, the ensemble approach, by incorporating multiple models in the final prediction, reduces the risk of overfitting the training data. Finally, the meta-learner NuSVC harnesses the strengths of each model, resulting in enhanced predictive performance. The key contributions of this work are outlined as follows:

- A Homogeneous stacked deep ensemble architecture, coined as HSTCN-NuSVC, is designed to capture hierarchical deep spatial-temporal features from input inertial signals without tedious signal pre-processing and the need for designing and selecting features in advance, boosting the model's generalisability.
- Stacking architecturally different TCN models allows the proposed HSTCN-NuSVC to perform multiscale feature extraction on the motion signals, enabling the extraction of low-to-high-level (hierarchical) deep features efficiently, as validated in subsequent ablation studies.
- In the proposed HSTCN-NuSVC, each base-level model, i.e., the TCN model, has fewer short filters, dilations and fewer dilated residual blocks to reduce the overall trainable parameters, lowering the model complexity and computational cost.
- Since retaining longer-term dependency is significant in time series classification problems, the proposed model is integrated with varying dilations and residual connections to prevent the gradients from exploding and vanishing during training.
- Extensive experimental analysis is conducted using the publicly accessible smartphone-based HAR dataset, UCI HAR. The experiments obey the user-independent protocol. The training and testing sets are mutually exclusive, and neither set shares samples from the same subject.



**Figure 1. Proposed HSTCN-NuSVC**



## 2- Homogeneous Stacked Temporal Convolutional Network with Nu-Support Vector Classifier (HSTCN-NuSVC)

Ensemble learning has become a go-to technique in classification problems regardless of the input data type. This technique combines multiple base learners, which may be susceptible to high bias or variance, and trains them to solve the same classification task. Researchers generally resort to ensemble learning algorithms when their primary focus is achieving high-performing predictive models. Ensemble learning algorithms can be classified into homogeneous and heterogeneous ensemble models. Homogeneous ensemble utilises the same learning algorithm for building level-zero/base models, whereas heterogeneous ensemble employs diverse learning algorithms to construct base learners [30]. The base models of an ensemble learning algorithm can be either a combination of traditional machine learning models, a combination of deep learning models, or both.

Bootstrap aggregation (bagging), boosting, and stacking are popular techniques in ensemble learning algorithms. In bagging, the dataset is sampled with replacement to split into equal-sized subsets, and several base models are trained independently with these subsets. Finally, the predictions are aggregated. The drawbacks of the bagging technique are that the ensemble model has a high chance of losing model interpretability due to high bias, and the overall training procedure can be computationally expensive [8, 9]. Boosting, unlike other techniques, is a sequential process where each subsequent model aims to correct the errors of the previous ones. This approach places greater emphasis on misclassified samples by assigning them higher weights, ensuring that future models prioritise these samples during training. However, a significant drawback of boosting is its sensitivity to outliers. This sensitivity arises because each new model in the sequence is focused on correcting the mistakes of the earlier models, allowing outliers to disproportionately influence the training process. Additionally, boosted models tend to be computationally intensive and may suffer from overfitting [9].

Unlike bagging and boosting algorithms, stacking involves combining predictions from multiple base learners to train a meta-model for final predictions. This technique reduces the ensemble model's susceptibility to bias and variance, enhancing generalisability during training and improving predictive performance. By organising base learners in a parallel structure, the model can capture diverse and unique features from motion signals, extracting more meaningful information about activity classes. In this study, a homogenous stacking model is developed, using identical deep learning algorithms as level-zero models for effective human activity recognition. Specifically, variants of TCN architectures serve as base learners, while a NuSVC classifier is employed as the meta-learner for the final prediction, as shown in Figure 1.

### 2-1-Base Models

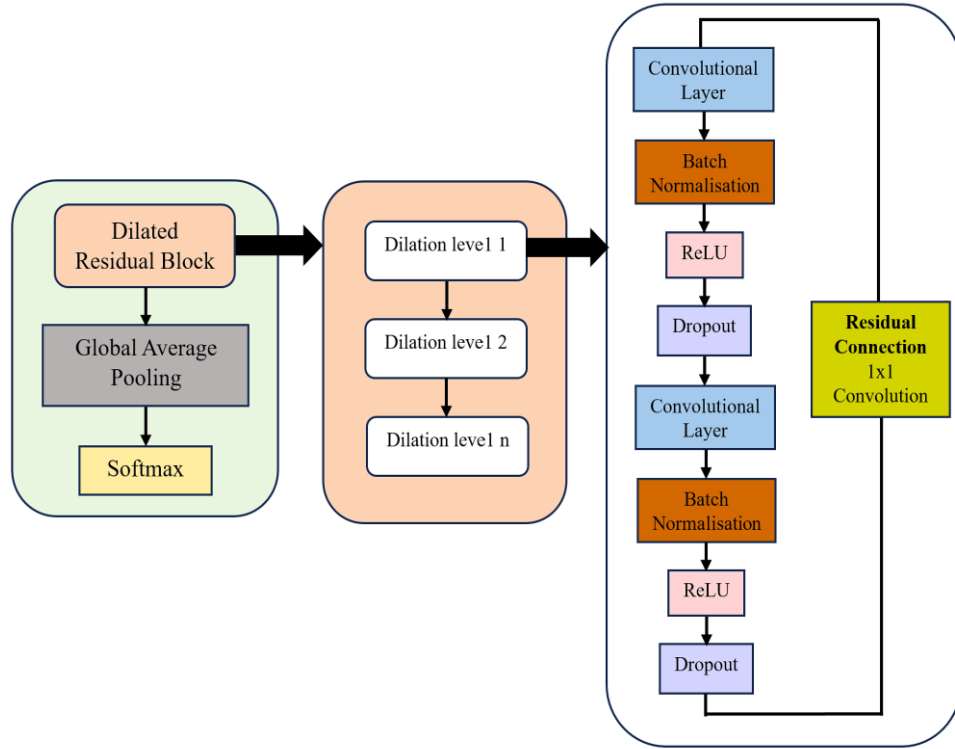
The proposed ensemble learner uses TCN variants as the base models. As illustrated in Figure 2, the TCN model has at least one dilated residual block, each with multiple dilation levels according to the dilation rate,  $d$ . Two one-dimensional (1D) dilated convolutional layers are arranged sequentially, and each has one batch normalisation, one activation, and one dropout layer at every dilation level. A residual connection consisting of a one-by-one convolutional layer is placed parallel to the 1D dilated convolutions. The outputs from this connection and 1D dilated convolution are concatenated and fed into the subsequent dilation level. The extracted deep spatial-temporal features are passed to the global average pooling (GAP) layer before being classified by a softmax classifier fitted with categorical cross-entropy loss. Each base model captures distinct underlying patterns from the input signals as the architectures and parameter configurations vary. The base models' architectures and parameter settings are chosen and optimised through a trial-and-error process. Specifically, we built and tested almost fifteen architectures by manipulating their hyperparameter configurations and selected only six high-performing TCN architectures as base models for our proposed model.

Integrating dilation into the proposed HSTCN-NuSVC ensemble is crucial in extracting extended time-dependent information from each input sample without raising the trainable parameters. In each base architecture (TCN model), the Dilated Residual Block implements 1D convolutional layers, which will be expanded according to the dilation rate settings, as illustrated in Figure 2. This practice could significantly reduce computational complexity. Applying dilations to convolutional kernels allows the model to expand the receptive fields and view a larger area of the input signal at a given time. To be specific, the convolutional kernel size is extended by adding zero between the convolutional kernel values, as shown in Figure 3. For instance, a receptive field of a convolutional kernel with a size of 5 and a dilation rate of 2 possesses the same coverage as the convolutional kernel with a size of 9 without any dilations. However, the former possesses lower trainable parameters. As the dilation level increases, the convolutional kernel becomes longer, improving its capability to capture longer temporal features from the input. This expanded field of view endows the model with the ability to garner richer information for classification. The equations of standard and dilated equations are as follows:

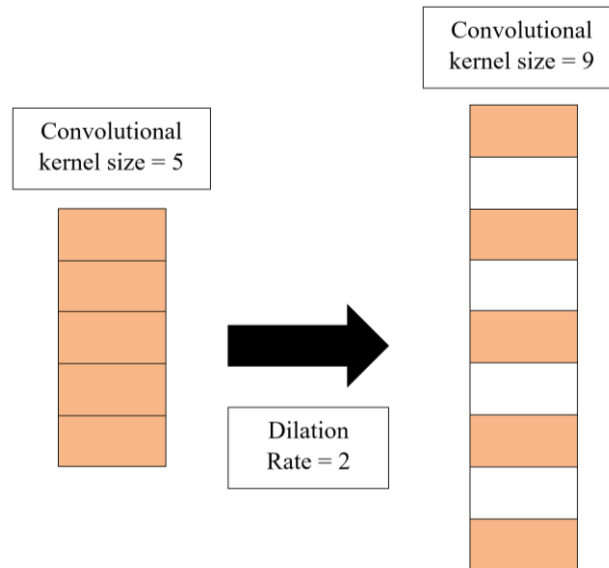
$$(x * k)(t) = \sum_{i=0} k(i) \cdot x(t - i) \quad (1)$$

$$(x_{*d} k)(t) = \sum_{i=0} k(i) \cdot x(t - d \cdot i) \quad (2)$$

where  $(x * k)(t)$  and  $(x_{*d} k)(t)$  are the output signal,  $k(i)$  is the convolutional kernel,  $t$  represents the position in the output sequence,  $i$  represents the position in the kernel, and  $d$  is the dilation rate. Equation 2 will be a standard convolution if the dilation rate is set to one. Since the proposed HSTCN-NuSVC contains multiple base TCN architectures, the dilation settings may differ from one model to another. The dilation rates used in our proposed model are 1, 2, 4 and 8. The dilation rate for each base TCN model is provided in the subsection Model Implementation.



**Figure 2. Architecture of base learner (TCN model)**



**Figure 3. 1D dilated convolution**

As mentioned, three components (i.e., batch normalisation, activation, and dropout) are added after each 1D dilated convolutional layer to enhance the model's performance during training. Batch normalisation is proposed by Ioffe & Szegedy [10] in order to minimise the internal covariate shift in CNN architectures during model training. In the training phase, the continual variation in the parameters of preceding layers causes the distributions within layers to fluctuate, slowing the model's convergence. To address this, batch normalisation is implemented, which standardises the input across small batches and linearly transforms them, ensuring a consistent mean and variance in each batch. In the proposed HSTCN-NuSVC system, ReLU activation is selected over alternatives such as sigmoid and tanh due to its non-saturating nature, computational efficiency, and resilience against vanishing gradient issues. Furthermore, overfitting, a common

issue that can degrade model performance, is effectively mitigated through dropout regularisation. This technique temporarily deactivates random neurones within hidden layers, preventing all neurone weights from being updated simultaneously. By doing so, it helps decorrelate neurone weights during training, promoting network sparsity.

Maintaining long-term dependencies within the network is crucial for time series classification, particularly in the HAR domain, as it prevents the loss of temporal information. However, as deep learning models grow more complex and deeper, they are increasingly prone to gradient vanishing and exploding problems. As training progresses, gradients can either diminish to negligible levels or increase excessively, causing the learning process to slow down or even halt. To counteract this, the proposed HSTCN-NuSVC employs residual connections, allowing activation functions to be skipped at each dilation level, thus addressing these challenges, as illustrated in Figure 2, to maintain the gradient stability during model training, boosting the classification performance. Generally, a residual connection is a pathway to pass extracted features to the subsequent layers without passing through any activation. Some existing works have implemented a one-by-one convolutional layer, a batch normalisation layer, or both. In our proposed model, a one-by-one convolutional layer is integrated into each residual connection between dilation levels in the Dilated Residual Block to reduce the feature dimensionality.

Lastly, the extracted deep spatial-temporal features from the dilated residual blocks are passed into the GAP layer to generate confidence maps. GAP prunes the overall trainable parameters in the proposed network, reducing the chances of model overfitting. Unlike standard fully connected dense layers, GAP does not require parameter optimisation, and its input does not need to be flattened. The averaged confidence maps are classified using a SoftMax classifier. This classifier computes the probability for each activity class and outputs a value ranging between zero and one. The total probability always sums up to one. We can identify the predicted class for a given sample by looking for the highest probability. SoftMax activation,  $\sigma()$ , can be written as follows:

$$\sigma(\vec{z}) = \frac{e^{z_i}}{\sum_{j=1}^N e^{z_j}} \quad (3)$$

where  $\vec{z}$  is the input vector,  $e^{z_i}$  is the exponential of the  $i^{th}$  element of  $\vec{z}$ ,  $N$  is the total number of activity classes, and  $\sum_{j=1}^N e^{z_j}$  is the sum of the exponentiated values of elements in  $\vec{z}$ . Categorical cross-entropy loss is equipped with this classifier to optimise the learning process during training. This loss determines the error between the true and target values to backpropagate them throughout the network until the minimum value is reached. This work considers different TCN architectures by manipulating the number of filters, filter sizes, type of padding, dilation rate, and number of dilated residual blocks. A detailed ablation study will be presented in the Architectural Study section.

## 2-2- Meta-Learner Model

The meta-learner model is crucial in stacking ensemble learning algorithms, as it utilises the predictions from base models to reduce bias and enhance the accuracy of the final classification. In this study, the diverse outputs from the base models (specifically, TCN architectures) are combined to form input data for training the meta-learner. This generated data functions as input features, while the ground truth values from the original dataset are used as the target classes for the meta-learner. The meta-learner is trained to optimally integrate the base models' predictions for a more accurate outcome. Although the predictions from the TCN base models capture complex data patterns, the relationship between those predictions and the final classification decision may still exhibit nonlinearity. Therefore, this work employs the NuSVC classifier, a Support Vector Machine (SVM) variant, as the meta-learner to combine predictions and effectively model nonlinear relationships using its kernel trick. The equation for NuSVC is as follows:

$$\min_{w, b, \xi, \rho} \left( \frac{1}{2} \|w\|^2 + \frac{1}{vN} \sum_{i=1}^N \xi_i - \rho \right) \quad (4)$$

subject to constraints:

$$y_i(w \cdot x_i + b) \geq \rho - \xi_i, \quad i = 1, \dots, N$$

$$\xi_i \geq 0, \quad i = 1, \dots, N$$

$$\sum_{i=1}^n \max(0, \rho - y_i(w \cdot x_i + b)) \leq vN$$

where  $x_i$  is the input data,  $y_i$  is the corresponding class label,  $w$  is the weight vector,  $v$  is the  $Nu$  parameter,  $b$  is the bias term,  $\xi_i$  is the slack variable,  $\rho$  is the parameter representing the margin distance, and  $N$  is the number of activity classes. Unlike the C-SVC classifier, where the  $C$  parameter ranges from zero to infinity, the  $v$  parameter only ranges between zero and one. So, it is easier to manage the complexity of the model as the number of support vectors can be adjusted by manipulating the  $Nu$  parameter. This classifier is good at handling noisy data and low interclass variance data. As a result, this classifier achieves better classification performance compared with the other traditional machine learning classifiers. The performance of the NuSVC classifier is evaluated and compared with existing popular machine learning classifiers, and the results are presented and discussed in the subsection Architectural Study.

### 3- Experimental Setup

A description of the adopted smartphone-based HAR dataset in this work is presented in the following subsection. Moreover, the proposed model's configuration and the performance measure used to evaluate the model's performance are also explained.

#### 3-1- Dataset Description

The proposed HSTCN-NuSVC model is trained and tested on a popular publicly available smartphone-based HAR dataset, the UCI HAR dataset, contributed to the UC Irvine Machine Learning Repository by Anguita et al. [11]. The dataset collected inertial signals from thirty volunteers aged 19 to 48 years. During the data collection, a smartphone with an embedded accelerometer and gyroscope was placed on a volunteer's waist. Each volunteer was required to complete six activities: walking, sitting, standing, walking upstairs, walking downstairs, and lying. Triaxial linear acceleration and triaxial angular velocity were collected from the embedded accelerometer and gyroscope at a constant rate of 50 Hz.

In this study, we followed a subject-independent evaluation protocol. Samples from one subset of volunteers were used for training, while samples from a separate group were reserved for testing. The participants in these two sets did not overlap, meaning no individual's data appeared in both the training and test sets. This approach is well-suited for real-time HAR applications, as it allows the model to predict activities for new users with little to no recalibration. To be specific, there will be variation in inertial signals between users when performing any activity and applying subject independence prevents the classifiers from overfitting to the training set and generalising better for new samples. The full dataset is split into two sets: the training set comprises samples from 70% of the volunteers, and the remaining subjects' samples are used for the test set. The inertial signals are pre-processed to remove any null values or noises. Then, pre-processed signals are partitioned into equal-sized time segments (128 readings/segment) using the sliding window technique. Moreover, triaxial body acceleration is also computed using the Butterworth low-pass filter to remove the gravitational component. Finally, the triaxial linear acceleration, body acceleration, and angular velocity are stacked to generate the desired input signal shape (128,9). The reason behind implementing the above-mentioned train-test split and pre-processing settings is to standardise the training and testing protocol, ensuring a fair comparison with existing HAR models.

#### 3-2- Model Implementation

This study's model development and experiments are carried out on a desktop with Intel® Core™ i9-12900K CPU with 2.20 GHz, 32 GB RAM, NVIDIA GeForce RTX 3080Ti, and 12 GB memory. Table 1 presents the hyperparameter configurations for each TCN architecture in the proposed HSTCN-NuSVC model. As mentioned, the optimal values for specific hyperparameters, such as batch size, number of dilated residual blocks, number of filters, filter sizes, dilation rates and padding, are identified through a trial-and-error process. The hyperparameter optimisation allows the deep learning models to reach their full potential. Moreover, the base TCN architectures are trained with dynamic learning rates by implementing the Reduce Learning Rate on the Plateau function to enhance learning. This function automatically reduces the learning rate by a preset percentage if the model's loss remains constant or rises. Besides the base learners, we also conducted parameter optimisation for the meta-learner (which will be discussed in the Architectural Study subsection). Table 2 records the finalised parameter settings of the proposed model's meta-learner. Additionally, we construct two baseline TCN models based on the original TCN architecture proposed by Lea et al. [31] and Bai et al. [32] in this study for performance comparison with our proposed HSTCN-NuSVC. This allows us to evaluate the effectiveness of our proposed model against these benchmarked TCN architectures. The architectural parameters of baseline TCN 1 [12] and baseline TCN 2 [13] are presented in Table 3.

**Table 1. Parameter settings of the base learner models**

Parameters	TCN 1	TCN 2	TCN 3	TCN 4	TCN 5	TCN 6
Batch size	32	32	32	32	32	32
Number of dilated residual blocks	1	2	2	1	2	2
Number of filters	32	40	40	64	40	50
Filter size	4	4	3	3	4	4
Dilation rate	1,2,4	1,2,4	1,2,4,8	1,2,4	1,2,4,8	1,2,4
Padding	causal	same	same	same	same	same
Dropout rate	0.05	0.05	0.05	0.05	0.05	0.05
Initial learning rate	0.01	0.01	0.01	0.01	0.01	0.01
Number of epochs	100	100	100	100	100	100
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam
Number of trainable parameters	24934	82366	87446	73862	111806	127456



**Table 2. Parameter configuration of the meta-learner model**

Parameters	Values
Nu	0.1
Gamma	1
Kernel	Linear
Hyperparameter tuning	Grid search cross-validation

**Table 3. Parameter settings of Baseline TCNs**

Parameters	Baseline TCN 1	Baseline TCN 2
Batch size	32	32
Number of TCN blocks	4	3
Number of filters	64	64
Filter size	3	3
Dilation rate	1,2,4,8	1,2,4,8
Padding	causal	causal
Dropout type and rate	Spatial dropout 0.05	Dropout 0.05
Activation function	Wavenet	ReLU
Normalisation	Channel	Weight
Global Average Pooling	No	No
Initial learning rate	0.01	0.01
Number of epochs	100	100
Optimizer	Adam	Adam
Number of trainable parameters	788742	1029598

### 3-3-Performance Measure

Several evaluation metrics, such as confusion matrix, precision, recall, accuracy and F1 score, are selected to evaluate the proposed model's performance and to perform performance comparison with other state-of-the-art models. The confusion matrix provides a comprehensive summary of the model's performance in such a way that the ground truth and the estimated activity class are presented in this single matrix. In other words, it offers the visualisation of four possible outcomes (e.g., true positives, true negatives, false positives and false negatives), as illustrated in Figure 4.

		Ground Truth	
		True	False
Prediction	True	True Positive (TP)	False Positive (FP)
	False	False Negative (FN)	True Negative (TN)

**Figure 4. Confusion Matrix**

Precision is an evaluation metric that finds all the dataset's relevant instances. The formula for precision is as follows:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (5)$$

Recall is a measure to identify all the relevant instances of a class in the dataset. Recall is formulated as:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (6)$$

Accuracy is a measure of computing the percentage of correct classification across all classes in the dataset. This metric is calculated as follows:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (7)$$

Lastly, F1 score is a measure that combines precision and recall into one evaluation metric. To be specific, this metric is also known as the harmonic mean of precision and recall. F1 score is computed using the following equation:

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (8)$$

## 4- Results and Discussion

The architectural study of the proposed HSTCN-NuSVC, as well as the analysis of the experimental results, is discussed in the following subsections. Furthermore, the performance of the proposed model is evaluated and compared with the other state-of-the-art models.

### 4-1-Architectural Study

The recognition performance of a learning model is highly contingent on its architectural parameter configuration. Hence, optimising the proposed HSTCN-NuSVC is crucial to enhance classification performance. In this section, an architectural study is conducted on the proposed model to examine the effect of architectural parameter variations. Several experiments are carried out by manipulating the number of base models, types of meta-learner models, and hyperparameters of the NuSVC classifier.

This study involves the development of eight TCN architectures, each with different parameter settings, which are used as base learners to examine the impact of varying the number of TCN models on the performance of the HSTCN-NuSVC. The experiment is conducted six times, with the number of base learners incrementally increasing by one in each iteration. Results are detailed in Table 4. The findings reveal that as the number of TCN models integrated into the proposed system grows from 3 to 6, the overall classification performance improves. This suggests that using multiple TCN models with varied architectures enables the system to capture a wider range of features from the input signals at different scales, enhancing the spatial-temporal information available for classification. However, after adding six or more TCN models, performance begins to decline. This reduction may be attributed to feature redundancy and potential overfitting of the training dataset.

**Table 4. The effects of the number of base learners on the classification performance**

Number of TCN Models	Total Parameters	Precision	Recall	F1 score	Accuracy (%)
3	194746	0.9596	0.9580	0.9585	95.72
4	268608	0.9498	0.9338	0.9374	93.42
5	380414	0.9711	0.9693	0.9695	96.91
6	507870	0.9748	0.9728	0.9731	97.25
7	581732	0.9615	0.9512	0.9540	95.15
8	611538	0.9726	0.9696	0.9699	96.95

Next, an experiment is conducted to identify an appropriate machine learning classifier for the meta-learner model. Seven different models, namely SVM, Multilayer Perceptron (MLP), KNN, Random Forest (RF), CatBoost, AdaBoost, and Nu-SVC, are considered in this experiment. Table 5 records the model performances. The CatBoost algorithm shows the worst performance among the other models. Besides that, it demands a longer time for model convergence. SVM, MLP, and KNN models perform satisfactorily by achieving F1 scores of 0.9709, 0.9714, and 0.9717, respectively, with reasonable training times. On the other hand, NuSVC exhibits slightly superior classification performance compared to the other machine learning models, achieving an F1 score of 0.9731. The superiority of NuSVC as a meta-learner can be attributed to its robustness to outliers, enabling it to generalise effectively on those unseen test data. Hence, the NuSVC classifier was selected as the meta-learner for the proposed HSTCN-NuSVC as it exhibited the best classification accuracy and F1 score among other models.

**Table 5. The effects of different types of meta-learner models on the classification performance**

Type of Meta-Learner Models	Precision	Recall	F1 score	Accuracy (%)
SVM	0.9730	0.9706	0.9709	97.05
MLP	0.9739	0.9711	0.9714	97.08
CatBoost	0.9576	0.9552	0.9548	95.52
KNN	0.9742	0.9714	0.9717	97.12
AdaBoost	0.9674	0.9635	0.9640	96.34
RF	0.9730	0.9704	0.9707	97.01
NuSVC	0.9748	0.9728	0.9731	97.25

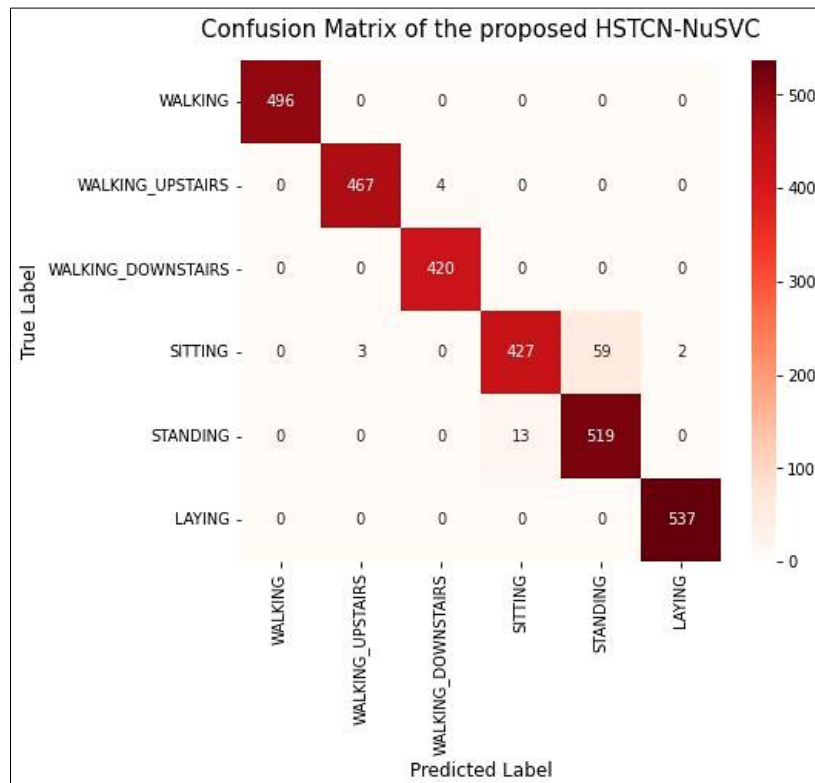
Lastly, hyperparameter tuning is conducted on the NuSVC classifier to optimise the recognition performance. Two primary hyperparameters, nu value and kernel type, are examined to assess their effect on the overall classification performance. The results of hyperparameter tuning are recorded in Table 6. From the empirical results, we can observe that the proposed HSTCN-NuSVC attains an optimal performance when the Nu value is set to 0.1. When the Nu value is increased from 0.1, the performance begins to drop and reaches a plateau. As for the kernel type of the NuSVC classifier, the performance differences are more significant. Four kernel options exist: linear, polynomial, sigmoid and radial basis functions. The empirical results demonstrate that the proposed model is able to achieve the best recognition performance when the kernel is set to the linear option. In contrast, the radial basis functions kernel shows the lowest accuracy compared to other kernels.

**Table 6. Hyperparameter tuning on NuSVC**

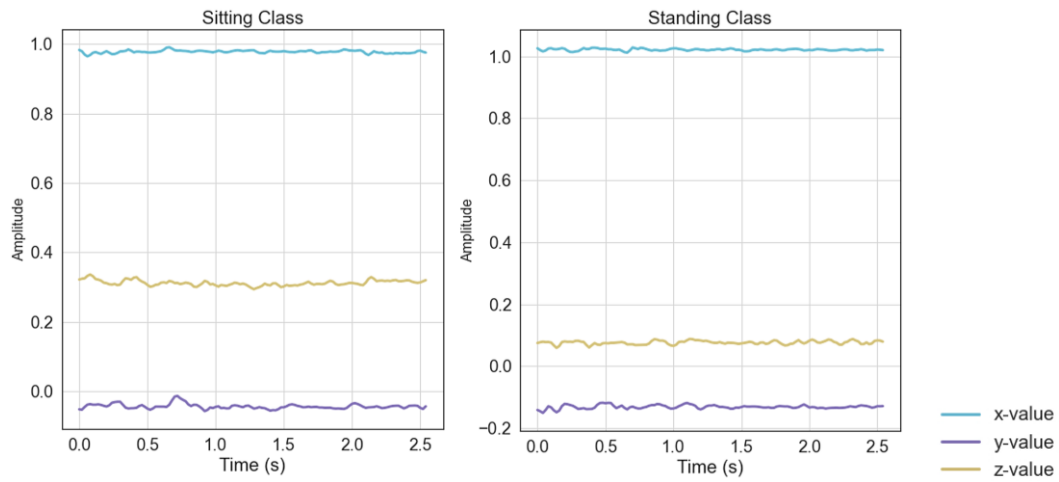
Hyperparameters of NuSVC	Hyperparameter Values	Precision	Recall	F1 score	Accuracy (%)
Nu	0.01	0.6921	0.6981	0.6948	68.48
	0.1	0.9748	0.9728	0.9731	97.25
	0.2	0.9740	0.9722	0.9724	97.18
	0.3	0.9739	0.9722	0.9724	97.18
Kernel	Linear	0.9748	0.9728	0.9731	97.25
	Polynomial	0.9613	0.9511	0.9540	95.15
	Sigmoid	0.9691	0.9657	0.9660	96.54
	Radial Basis Function	0.9597	0.9474	0.9508	94.81

#### 4-2- Performance Comparison with State-Of-The-Art Models

The proposed HSTCN-NuSVC is evaluated on the test set to assess the model's efficacy on unseen data. Figure 5 illustrates the confusion matrix of the proposed model. The resulting matrix shows that the proposed HSTCN-NuSVC possesses relatively low misclassifications across all data. The *sitting* class has slightly more misclassification than the other activity classes. Fifty-nine samples of the *sitting* class are misclassified as the *standing* class. This misclassification might be due to the low interclass variance between the *sitting* and *standing* classes, as illustrated in Figure 6. From the figure, we can see that certain common patterns between the signals of both classes make it challenging for the model to discriminate between them.

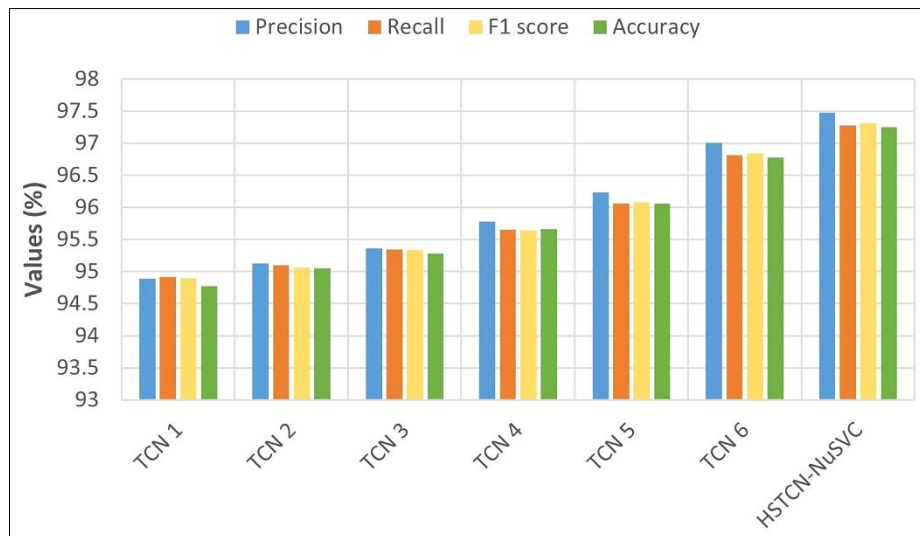


**Figure 5. Confusion Matrix of HSTCN-NuSVC model**



**Figure 6.** Input signal samples of sitting and standing classes

Furthermore, an overall classification performance comparison between the proposed HSTCN-NuSVC and the six individual base learners is illustrated in Figure 7. The proposed HSTCN-NuSVC dominates the base learner models across all the evaluation metrics. This indicates that the staking ensemble learning algorithm enables the proposed model to perform diverse and multiscale feature extraction on input inertial signals, gaining more salient deep features that distinguish between activities well. Additionally, the NuSVC classifier enhances the performance further by integrating the base learners' predictions. The classifier leverages and synthesises the diverse analytical decisions of these base learners to produce a more accurate final prediction decision. The performance of the proposed HSTCN-NuSVC in terms of accuracy, model parameters, training time, and inference time is recorded in Table 7. Our proposed model attained a high accuracy of 97.25% with only 0.51 million parameters. The training of the proposed ensemble took about 42 minutes. However, our model has a short inference time of 2.6s to make predictions on 2947 samples.



**Figure 7.** Performance comparison between the proposed HSTCN-NuSVC and the individual base learner models

**Table 7.** Performance of our proposed HSTCN-NuSVC

Criteria	Values
Accuracy (%)	97.25
Model parameters (million)	0.51
Training time (min)	42
Inference time (s)	2.6

Table 8 compares classification performance between HSTCN-NuSVC and other state-of-the-art models. These include handcrafted feature-based, deep learning and ensemble learning models. From the findings, we can observe that the proposed HSTCN-NuSVC outperforms most of the existing methods regardless of their types. Anguita et al. [11] published the UCI HAR dataset and performed benchmark analysis using popular machine learning classifiers. The

authors found that the SVM classifier had the highest accuracy, 96%, compared to the other models. Our proposed model achieves approximately 1.25% accuracy compared to the benchmark performance without requiring manual feature engineering/selections. RK-KNN [20] had the second-highest accuracy among handcrafted feature-based methods and a moderate inference time of 44 ms/sample. Our proposed model surpasses the handcrafted feature-based models in recognition accuracy and inference time due to its automatic feature learning capability and generalisability on unseen data. In other words, our proposed model does not require time-consuming manual feature modelling, ranking and selection, which is vulnerable to bias. This increases the models' chances of losing implicit temporal patterns from motion signals, causing the performance to deteriorate.

**Table 8. Performance comparison with the state-of-the-art methods**

Methods	Accuracy (%)	Model parameter (million)	Inference time (ms/sample)
Support Vector Machine (SVM) [11]	96.00*	-	-
K-Nearest Neighbours (KNN) [31]	90.46*	-	-
RK-KNN [20]	92.67*	-	44
Hierarchical multi-view aggregation network (HVMAN) [24]	94.70*	-	-
Multilayer Perceptron model (MLP) [33]	95.00*	-	-
Attention induced multi-head CNN [22]	95.32*	1.51*	-
Layer-Wise Training CNN with Smaller Filters [23]	96.90*	-	110
Multichannel CNN [24]	95.25*	-	-
ResNet + SelectConv [34]	97.28*	0.84*	-
Deep Residual Bidirectional LSTM [28]	93.60*	-	-
Bidirectional LSTM [29]	93.79*	-	-
Stacked LSTM [35]	93.13*	-	-
InnoHAR [36]	94.50*	-	65.09
CNN-LSTM [37]	92.13*	-	-
Multi-input CNN-GRU [38]	96.20*	-	-
Dilated TCN [12]	93.80*	0.15*	-
Encoder-Decoder TCN [18]	94.60*	0.16*	-
MSTCN [19]	97.42*	3.75*	3.3
Light-MHTCN [39]	96.48*	0.21*	2.2
Baseline TCN 1	90.87	1.73	2.7
Baseline TCN 2	92.98	0.28	2.4
Cascade Ensemble Learning (CELearning) [40]	96.67*	-	-
Ensem-HAR [29]	95.05*	-	-
Ensemble Learning Algorithm (ELA) [28]	96.70*	-	-
Stacked Generalisation with Wrapper-Based Feature Selection [41]	97.01*	-	-
Multi-input Hybrid CNN [27]	94.00*	-	-
ResLSTM [42]	96.34*	0.58*	-
Hybrid CNN-SVM [43]	96.13*	-	-
Hybrid CNN and Stacked LSTM [26]	93.40*	0.46*	-
WISNet [21]	95.66*	0.70*	-
Proposed HSTCN-NuSVC	97.25	0.51	0.90

\* Results extracted from the respective articles

The backbone of attention-induced multi-head CNN [22], CNN with smaller filters [23], Multichannel CNN [24], ResNet+SelectConv [25], Multi-input Hybrid CNN [27] and WISNet [21] is CNN architecture. ResNet+SelectConv achieves slightly higher classification performance than the proposed HSTCN-NuSVC model with a marginal difference of 0.03%. However, our proposed architecture achieves 97.25% accuracy with only 0.51 million parameters, which is 330000 less trainable than ResNet+SelectConv. CNN with smaller filters [23] attains the second-best accuracy among the CNN models with an accuracy of 96.90%, but this model had the highest inference time of 110 ms/sample. The proposed HSTCN-NuSVC achieves high recognition accuracy with a very short inference time. Unlike the CNN



architectures, the base learner models of our proposed model are built based on TCN architecture, which excels at capturing longer time-dependent features from the input signals, making it a good solution for learning temporal sequences in human motion.

Although recurrent models are optimal for time series classification, the performance of these models is moderate compared to our proposed architecture. The classification accuracy for Deep Residual Bidirectional LSTM [28], Bidirectional LSTM [29], and Stacked LSTM [35] was around 93%, which is ~4% lower than the proposed HSTCN-NuSVC. Besides that, ResLSTM [44] achieved slightly higher accuracy than other recurrent networks with relatively low parameters (0.58 million). Nevertheless, the performance of our proposed architecture is better than ResLSTM, as our model attains slightly higher accuracy of ~1% and 70000 fewer parameters. Several authors developed hybrid models combining CNN and recurrent architectures, such as InnoHAR [36], CNN-LSTM [37], Multi-input CNN-GRU [38], Hybrid CNN-SVM [43] and Hybrid CNN and Stacked LSTM [26], to leverage the benefits of both architectures. However, the performance of the hybrid networks is suboptimal compared to the proposed HSTCN-NuSVC. Since the main constituent of the proposed architecture is TCN models, the model is able to retain longer-term dependencies by preserving a longer effective history of the motion signals, contributing to superior prediction performance compared to the recurrent models.

Besides that, there are several works based on ensemble learning, namely Cascade Ensemble Learning (CELearning) [40], Ensemble Learning Algorithm (ELA) [28], Ensem-HAR [29], and Stacked Generalisation with a Wrapper-Based Feature Selection [41]. From Table 8, we notice that Stacked Generalisation with a Wrapper-Based Feature Selection [43] achieves the second-highest accuracy of 97.01% among the ensemble learning models, following behind our proposed model. Nonetheless, this ensemble model relies on manual feature extraction techniques, rendering it labour-intensive. As previously mentioned, the base learners of our proposed architecture are based on TCN models. Hence, we compared the performance of the proposed HSTCN-NuSVC with other popular TCN variants, including Dilated TCN [18], Encoder-Decoder TCN [18], MSTCN [19], and Light-MHTCN [39], in the HAR domain. Our proposed model attains better accuracy than most TCN models, except the MSTCN model [19], since our model integrates multiple TCN models with varied structural characteristics. This design promotes the model to extract hierarchical features at different scales, providing exclusive insights into the underlying pattern of the motion signals. MSTCN [19] achieved higher accuracy than the proposed HSTCN-NuSVC with a minute difference of 0.17%. However, the overall performance of our proposed architecture is still considered superior to MSTCN, as our model has approximately seven times fewer model parameters and a shorter inference time than MSTCN. Additionally, we built two baseline TCN models according to the hyperparameter settings provided in Table 3 to compare them with the proposed HSTCN-NuSVC. These models are based on the works of Lea et al. [31] and Bai et al. [32]. Although both models had relatively low trainable parameters and inference time, they performed poorly during testing.

## 5- Conclusion

A homogenous stacked ensemble deep learning model, HSTCN-NuSVC, is proposed to perform human activity recognition without requiring manual feature modelling and selection. This model integrates multiple base learners along with a meta-learner component. The base learner models leverage the Temporal Convolutional Network (TCN) architecture, which excels in time series classification tasks. Key enhancements include the replacement of spatial dropout with standard dropout, channel and weight normalisation with batch normalisation, and the use of ReLU activation in place of WaveNet activation to improve base learner performance. By stacking several TCN models with varying kernel sizes, filter numbers, and dilation rates, the model effectively performs deep multiscale feature extraction, capturing essential information about human movement. Additionally, incorporating dilations in each convolutional layer expands the convolutional kernel's receptive field, allowing the model to process information across broader temporal spans, thereby enhancing its capacity to analyse intricate dependencies in human motion data. Importantly, this approach does not substantially increase the number of learnable parameters, thus avoiding significant computational overhead. Moreover, the use of short filters, GAP layers, and dilated residual blocks helps to reduce the overall parameter count, making the model more computationally efficient. To address issues related to vanishing and exploding gradients, residual connections are employed at each dilation level within the dilated residual blocks. The second phase of the HSTCN-NuSVC model involves the meta-learner phase. The NuSVC classifier is selected as the meta-learner model as it exhibits higher performance than other machine learning models. The meta-learner amalgamates the individual predictions from each base learner in an optimal manner to produce the final predictions. Finally, the proposed HSTCN-NuSVC is evaluated on the benchmark smartphone-based HAR database, i.e., UCI HAR, using the subject-independent protocol. The empirical results demonstrate that the proposed model outperforms most existing human activity recognition models with an accuracy of 97.25%, fewer learnable parameters, and a short inference time. Since this work only includes one smartphone-based HAR dataset, UCI HAR, consisting of six daily activities, we plan to utilise other smartphone-based HAR datasets with more diverse activities and volunteers to validate the performance of our proposed HSTCN-NuSVC and the generalising ability in our future work.

## 6- Declarations

### 6-1-Author Contributions

Conceptualization, S.R.S.; methodology, S.R.S.; software, S.R.S.; validation, S.R.S. and Y.H.P.; formal analysis, S.R.S.; investigation, S.R.S.; resources, S.R.S.; writing—original draft preparation, S.R.S.; writing—review and editing, S.R.S., Y.H.P., O.S.Y., and L.Z.Y.; visualization, S.R.S.; supervision, Y.H.P.; project administration, Y.H.P.; funding acquisition, Y.H.P. All authors have read and agreed to the published version of the manuscript.

### 6-2-Data Availability Statement

The data presented in this study are openly available in UC Irvine Machine Learning Repository at 10.24432/C54S4K.

### 6-3-Funding

This research is supported by Fundamental Research Grant Scheme (FRGS), FRGS/1/2020/ICT02/MMU/02/7. The authors fully acknowledged the Ministry of Higher Education Malaysia (MOHE) for the approved fund, which makes this important research viable and effective.

### 6-4-Institutional Review Board Statement

Not applicable.

### 6-5-Informed Consent Statement

Not applicable.

### 6-6-Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

## 7- References

- [1] Zhai, Y., Nasser, N., Pöttgen, J., Gezhelbash, E., Heesen, C., & Stellmann, J. P. (2020). Smartphone Accelerometry: A Smart and Reliable Measurement of Real-Life Physical Activity in Multiple Sclerosis and Healthy Individuals. *Frontiers in Neurology*, 11. doi:10.3389/fneur.2020.00688.
- [2] Dhiman, C., & Vishwakarma, D. K. (2019). A review of state-of-the-art techniques for abnormal human activity recognition. *Engineering Applications of Artificial Intelligence*, 77, 21–45. doi:10.1016/j.engappai.2018.08.014.
- [3] Li, K., Wu, J., Zhao, X., & Tan, M. (2019). Real-Time Human-Robot Interaction for a Service Robot Based on 3D Human Activity Recognition and Human-Mimicking Decision Mechanism. 8th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, CYBER 2018, 2018, 498–503. doi:10.1109/CYBER.2018.8688272.
- [4] danova, M., Voronin, V. V., Semenishchev, E., Ilyukhin, Y. V., & Zelensky, A. (2020). Human activity recognition for efficient human-robot collaboration. *Artificial Intelligence and Machine Learning in Defense Applications II*, 16. doi:10.1117/12.2574133.
- [5] Anagnostis, A., Benos, L., Tsaopoulos, D., Tagarakis, A., Tsolakis, N., & Bochtis, D. (2021). Human activity recognition through recurrent neural networks for human–robot interaction in agriculture. *Applied Sciences (Switzerland)*, 11(5), 2188. doi:10.3390/app11052188.
- [6] Tharwat, A., Mahdi, H., Elhoseny, M., & Hassanien, A. E. (2018). Recognizing human activity in mobile crowdsensing environment using optimized k-NN algorithm. *Expert Systems with Applications*, 107, 32–44. doi:10.1016/j.eswa.2018.04.017.
- [7] Batool, M., Jalal, A., & Kim, K. (2019). Sensors technologies for human activity analysis based on SVM optimized by PSO algorithm. *International Conference on Applied and Engineering Mathematics (ICAEM-2019)*, 145–150. doi:10.1109/ICAEM.2019.8853770.
- [8] Ansari, G., Ahmad, T., & Doja, M. N. (2019). Hybrid filter–wrapper feature selection method for sentiment classification. *Arabian Journal for Science and Engineering*, 44, 9191–9208. doi:10.1007/s13369-019-04064-6.
- [9] Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, 35(2), 757–774. doi:10.1016/j.jksuci.2023.01.014.
- [10] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. 32<sup>nd</sup> International Conference on Machine Learning, ICML 2015, 6–11 July, 2015, Lille, France.

- [11] Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013, April). A public domain dataset for human activity recognition using smartphones. ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 24-26 April, 2013, Bruges, Belgium.
- [12] Mohsen, S., Elkaseer, A., Scholz, S.G. (2022). Human Activity Recognition Using K-Nearest Neighbor Machine Learning Algorithm. Sustainable Design and Manufacturing. KES-SDM 2021. Smart Innovation, Systems and Technologies, 262, Springer, Singapore. doi:10.1007/978-981-16-6128-0\_29.
- [13] Kee, Y. J., Zainudin, M. S., Idris, M. I., Ramlee, R. H., & Kamarudin, M. R. (2020). Activity recognition on subject independent using machine learning. Cybernetics and Information Technologies, 20(3), 64-74. doi:10.2478/cait-2020-0028.
- [14] Hamad, R. A., Kimura, M., Yang, L., Woo, W. L., & Wei, B. (2021). Dilated causal convolution with multi-head self-attention for sensor human activity recognition. Neural Computing and Applications, 33(20), 13705–13722. doi:10.1007/s00521-021-06007-5.
- [15] Han, C., Zhang, L., Tang, Y., Huang, W., Min, F., & He, J. (2022). Human activity recognition using wearable sensors by heterogeneous convolutional neural networks. Expert Systems with Applications, 198, 116764. doi:10.1016/j.eswa.2022.116764.
- [16] Hernandez, F., Suarez, L. F., Villamizar, J., & Altuve, M. (2019). Human Activity Recognition on Smartphones Using a Bidirectional LSTM Network. 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), 1–5. doi:10.1109/stsiva.2019.8730249.
- [17] Pan, J., Hu, Z., Yin, S., & Li, M. (2022). GRU with Dual Attentions for Sensor-Based Human Activity Recognition. Electronics (Switzerland), 11(11). doi:10.3390/electronics11111797.
- [18] Nair, N., Thomas, C., & Jayagopi, D. B. (2018). Human Activity Recognition Using Temporal Convolutional Network. Proceedings of the 5<sup>th</sup> International Workshop on Sensor-Based Activity Recognition and Interaction, 1–8. doi:10.1145/3266157.3266221.
- [19] Raja Sekaran, S., Pang, Y. H., Ling, G. F., & Yin, O. S. (2022). MSTCN: A multiscale temporal convolutional network for user independent human activity recognition. F1000Research, 10, 1261. doi:10.12688/f1000research.73175.2.
- [20] Liu, Z., Li, S., Hao, J., Hu, J., & Pan, M. (2021). An Efficient and Fast Model Reduced Kernel KNN for Human Activity Recognition. Journal of Advanced Transportation, 2026895. doi:10.1155/2021/2026895.
- [21] Khan, Z. N., & Ahmad, J. (2021). Attention induced multi-head convolutional neural network for human activity recognition. Applied Soft Computing, 110, 107671. doi:10.1016/j.asoc.2021.107671.
- [22] Sharen, H., Jani Anbarasi, L., Rukmani, P., Gandomi, A. H., Neeraja, R., & Narendra, M. (2024). WISNet: A deep neural network based human activity recognition system. Expert Systems with Applications, 258, 124999. doi:10.1016/j.eswa.2024.124999.
- [23] Tang, Y., Teng, Q., Zhang, L., Min, F., & He, J. (2021). Layer-Wise Training Convolutional Neural Networks with Smaller Filters for Human Activity Recognition Using Wearable Sensors. IEEE Sensors Journal, 21(1), 581–592. doi:10.1109/JSEN.2020.3015521.
- [24] Zhang, X., Wong, Y., Kankanhalli, M. S., & Geng, W. (2019). Hierarchical multi-view aggregation network for sensor-based human activity recognition. PLoS ONE, 14(9). doi:10.1371/journal.pone.0221390.
- [25] Huang, W., Zhang, L., Gao, W., Min, F., & He, J. (2021). Shallow Convolutional Neural Networks for Human Activity Recognition Using Wearable Sensors. IEEE Transactions on Instrumentation and Measurement, 70. doi:10.1109/TIM.2021.3091990.
- [26] Dubey, A., & Zacharias, J. (2024). IMU Data Based HAR Using Hybrid Model of CNN & Stacked LSTM. International Conference on Advancements in Power, Communication and Intelligent Systems, APCI 2024, 2024, 1–6. doi:10.1109/APCI61480.2024.10616429.
- [27] Lai, Y. C., Kan, Y. C., Hsu, K. C., & Lin, H. C. (2024). Multiple inputs modeling of hybrid convolutional neural networks for human activity recognition. Biomedical Signal Processing and Control, 92, 106034. doi:10.1016/j.bspc.2024.106034.
- [28] Bhattacharya, D., Sharma, D., Kim, W., Ijaz, M. F., & Singh, P. K. (2022). Ensem-HAR: An Ensemble Deep Learning Model for Smartphone Sensor-Based Human Activity Recognition for Measurement of Elderly Health Monitoring. Biosensors, 12(6). doi:10.3390/bios12060393.
- [29] Tan, T. H., Wu, J. Y., Liu, S. H., & Gochoo, M. (2022). Human Activity Recognition Using an Ensemble Learning Algorithm with Smartphone Sensor Data. Electronics (Switzerland), 11(3), 1–17. doi:10.3390/electronics11030322.
- [30] Rooney, N., Patterson, D., Tsymbal, A., & Anand, S. (2004). Random subsampling for regression ensembles. Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, 12-14 May, 2004, Florida, United States.

- [31] Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2017). Temporal Convolutional Networks for Action Segmentation and Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, United States. doi:10.1109/cvpr.2017.113.
- [32] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint, arXiv:1803.01271. doi:10.48550/arXiv.1803.01271.
- [33] Ogbuabor, G., & La, R. (2018). Human Activity Recognition for Healthcare using Smartphones. Proceedings of the 2018 10th International Conference on Machine Learning and Computing, 41–46. doi:10.1145/3195106.3195157.
- [34] Huang, W., Zhang, L., Teng, Q., Song, C., & He, J. (2021). The Convolutional Neural Networks Training with Channel-Selectivity for Human Activity Recognition Based on Sensors. IEEE Journal of Biomedical and Health Informatics, 25(10), 3834–3843. doi:10.1109/JBHI.2021.3092396.
- [35] Ullah, M., Ullah, H., Khan, S. D., & Cheikh, F. A. (2019). Stacked LSTM Network for Human Activity Recognition Using Smartphone Data. Proceedings - European Workshop on Visual Information Processing, EUVIP, 2019-October, 175–180. doi:10.1109/EUVIP47703.2019.8946180.
- [36] Xu, C., Chai, D., He, J., Zhang, X., & Duan, S. (2019). InnoHAR: A deep neural network for complex human activity recognition. IEEE Access, 7, 9893–9902. doi:10.1109/ACCESS.2018.2890675.
- [37] Mutegeki, R., & Han, D. S. (2020). A CNN-LSTM Approach to Human Activity Recognition. 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIC), 362–366. doi:10.1109/icaic48513.2020.9065078.
- [38] Dua, N., Singh, S. N., & Semwal, V. B. (2021). Multi-input CNN-GRU based human activity recognition using wearable sensors. Computing, 103(7), 1461–1478. doi:10.1007/s00607-021-00928-8.
- [39] Raja Sekaran, S., Han, P. Y., & Yin, O. S. (2023). Smartphone-based human activity recognition using lightweight multiheaded temporal convolutional network. Expert Systems with Applications, 227, 120132. doi:10.1016/j.eswa.2023.120132.
- [40] Xu, S., Tang, Q., Jin, L., & Pan, Z. (2019). A cascade ensemble learning model for human activity recognition with smartphones. Sensors (Switzerland), 19(10), 1–17. doi:10.3390/s19102307.
- [41] Bhavan, A., & Aggarwal, S. (2018). Stacked Generalization with Wrapper-Based Feature Selection for Human Activity Recognition. 2018 IEEE Symposium Series on Computational Intelligence (SSCI), 1064–1068. doi:10.1109/ssci.2018.8628830.
- [42] AlMuhaideb, S., AlAbdulkarim, L., AlShahrani, D. M., AlDhubaib, H., & AlSadoun, D. E. (2024). Achieving More with Less: A Lightweight Deep Learning Solution for Advanced Human Activity Recognition (HAR). Sensors, 24(16), 5436. doi:10.3390/s24165436.
- [43] Charabi, I., Abidine, M. B., & Fergani, B. (2024). Sensor-Based Human Activity Recognition Using a Hybrid CNN-SVM Approach. 2024 8<sup>th</sup> International Conference on Image and Signal Processing and Their Applications (ISPA), 1–6. doi:10.1109/ispa59904.2024.10536787.