Hybrid Spatial-Temporal Deep Learning Architectures for FMCW Radar-Based Human Activity Recognition

Daffa Ahmadhan Khusumah[1*], Fiky Yosef Suratman[2], Hesty Susanti[3]
[1-3]Teknik Elektro, Fakultas Teknik Elektro, Telkom University, Indonesia

Corresponding Author's e-mail: *daffaahmadhan@student.telkomuniversity.ac.id*

ARMADA
JURNAL PENELITIAN MULTIDISIPLIN

e-ISSN: 2964-2981

ARMADA : Jurnal Penelitian Multidisiplin

https://ejournal.45mataram.ac.id/index.php/armada

Vol. 04, No. 07 Juli, 2026

Page: 2739-2750

DOI:

https://doi.org/10.55681/armada.v4i7.3001

Article History:

Received: Mei 01, 2026

Revised: Juni 13, 2026

Accepted: Juli 19, 2026

Abstract : Human Activity Recognition (HAR) supports intelligent healthcare, surveillance, assisted living, and human–machine interaction. Vision-based methods are often limited by privacy concerns, illumination changes, and occlusion. This study proposes a hybrid spatial–temporal deep learning framework for FMCW radar-based HAR using micro-Doppler spectrograms. Four architectures are compared: 3D CNN–LSTM, 3D Bi-LSTM–CNN, CNN–Dilated Convolution–LSTM, and a Hybrid Ensemble CNN-LSTM with a Decision Tree classifier. Radar processing includes beat-frequency extraction, Range FFT, Doppler FFT, clutter suppression, and spectrogram generation. Convolutional layers extract spatial features, while LSTM and Bi-LSTM networks model temporal dependencies; dilated convolution expands the receptive field efficiently. Experimental results show that the hybrid models outperform conventional CNN and standalone LSTM approaches in accuracy, robustness, and generalisation. The hybrid ensemble achieves the best performance by combining spatial–temporal learning with ensemble optimisation while remaining effective in noisy environments and preserving user privacy.

Keywords : FMCW Radar, Human Activity Recognition, CNN, Bi-LSTM, Dilated Convolution, Spatial–Temporal Learn-ing

Abstrak : *Pengenalan Aktivitas Manusia (Human Activity Recognition atau HAR) mendukung layanan kesehatan cerdas, pengawasan, kehidupan berbantuan, dan interaksi manusia mesin. Metode berbasis visi sering dibatasi oleh masalah privasi, perubahan pencahayaan, dan oklusi. Penelitian ini mengusulkan kerangka pembelajaran mendalam hibrida spasial–temporal untuk HAR berbasis radar FMCW menggunakan spektrogram mikro-Doppler. Empat arsitektur dibandingkan, yaitu 3D CNN–LSTM, 3D Bi-LSTM–CNN, CNN–Dilated Convolution–LSTM, dan Hybrid Ensemble CNN-LSTM dengan klasifikator Decision Tree. Pemrosesan radar mencakup ekstraksi frekuensi beat, Range FFT, Doppler FFT, penekanan clutter, dan pembentukan spektrogram. Lapisan konvolusional mengekstraksi fitur spasial, sedangkan jaringan LSTM dan Bi-LSTM memodelkan dependensi temporal; dilated convolution memperluas medan reseptif secara efisien. Hasil eksperimen menunjukkan bahwa model hibrida mengungguli CNN konvensional dan LSTM mandiri dalam akurasi, ketahanan, dan generalisasi. Model hybrid ensemble menghasilkan kinerja terbaik dengan menggabungkan pembelajaran spasial–temporal dan optimasi ensembel,*

sekaligus tetap efektif dalam lingkungan bising serta menjaga privasi pengguna.

Kata Kunci: *Radar FMCW, Pengenalan Aktivitas Manusia, CNN, Bi-LSTM, Konvolusi Berdilatasi, Pembelajaran Spasial–Temporal*

INTRODUCTION

Human Activity Recognition (HAR) is an important component of intelligent sensing systems, enabling human movements and behaviours to be identified automatically from sensor data. Tan et al. (2024) highlighted the growing relevance of HAR in healthcare monitoring, elderly assistance, smart homes, rehabilitation, security surveillance, and human–machine interaction. In these applications, HAR systems are expected not only to identify routine activities but also to detect abnormal events and support timely responses. Conventional HAR systems commonly rely on RGB cameras, infrared cameras, depth sensors, or wearable devices. Although vision-based systems provide detailed spatial information, Zhang et al. (2024) noted that their performance is affected by illumination changes, occlusion, shadows, and viewing angles. Camera-based monitoring may also expose facial and physical information, creating privacy concerns in homes and healthcare environments. Diraco et al. (2025) therefore emphasised the need for non-visual sensing technologies that can monitor human activity without recording identifiable images. Wearable sensors offer another alternative, but their effectiveness depends on users consistently wearing and correctly positioning the devices.

Frequency-Modulated Continuous-Wave (FMCW) radar provides a promising solution because it captures human motion through reflected electromagnetic waves rather than visual appearance. FMCW radar transmits frequency-modulated chirp signals and analyses the frequency difference between transmitted and received signals to estimate target range and velocity. Movements of the arms, legs, and torso produce additional frequency modulations known as micro-Doppler signatures. Tan et al. (2024) demonstrated that these signatures can be transformed into time–frequency spectrograms using the Short-Time Fourier Transform, allowing different human activities to be represented without revealing the user's identity. Despite these advantages, FMCW radar-based HAR remains challenging. Activities involving similar body movements may produce overlapping micro-Doppler patterns, making them difficult to distinguish. Wu et al. (2025) reported that similar activities require more representative spatial and temporal features because their radar signatures often share comparable time–frequency characteristics. Radar signals may also be affected by thermal noise, static clutter, multipath reflections, body orientation, movement speed, and subject-to-radar distance. Chen et al. (2025) found that environmental changes and multipath effects can reduce the generalisation performance of radar-based recognition models.

Deep learning enables discriminative radar features to be learned automatically rather than relying entirely on manually designed features. Convolutional Neural Networks (CNNs) are effective in extracting local and hierarchical spatial patterns from micro-Doppler spectrograms. Zhang et al. (2024) showed that enhanced convolutional residual structures can improve the extraction of human-motion features from FMCW radar data. However, conventional CNNs mainly focus on spatial information and have limited ability to preserve long-term movement sequences. Long Short-Term Memory (LSTM) networks address this limitation by modelling temporal dependencies through gated memory mechanisms. Wu et al. (2025) combined CNN and Bidirectional LSTM (Bi-LSTM) to connect high-level spatial features with temporal changes in human activities. Bi-LSTM processes information in both forward and backward directions, allowing the model to consider preceding and subsequent movement patterns simultaneously. Lin et al. (2024) demonstrated the value of CNN-BiLSTM for combining range, azimuth, and temporal information in multi-activity classification.

Another limitation concerns the restricted receptive field of standard convolution. Dilated convolution expands the receptive field by introducing gaps between kernel elements, enabling wider temporal contexts to be learned without a proportional increase in model parameters. Kakuba et al. (2024) applied dilated causal convolution to radar point-cloud data and showed its potential for modelling broader temporal relationships. Combining dilated convolution with

recurrent learning may therefore improve the recognition of long-duration and multi-scale activity patterns. Recent studies represent the current state of the art in radar-based HAR. Tan et al. (2024) combined dual-stream 1D-CNN and BiGRU networks to process magnitude and phase information. Lin et al. (2024) integrated statistical, time–frequency, range–azimuth–time, CNN-BiLSTM, PCANet, and random forest features for multi-activity classification. Zhang et al. (2024) developed an asymmetric convolutional residual architecture, while Wu et al. (2025) combined CNN-BiLSTM with class activation mapping to distinguish similar activities. However, these studies generally concentrated on a single main architecture and used different datasets, activity classes, and evaluation procedures, making direct performance comparisons difficult.

A research gap therefore remains in the systematic comparison of spatial–temporal architectures under the same FMCW radar processing pipeline and experimental conditions. Existing studies have not sufficiently compared three-dimensional convolution, bidirectional temporal modelling, dilated convolution, and ensemble classification within one framework. Furthermore, improvements in accuracy are often accompanied by increased model size, inference latency, and computational requirements. Lim et al. (2024) stressed that practical radar-based HAR systems must balance recognition performance with computational efficiency. To address these limitations, this study proposes and compares four hybrid spatial–temporal architectures: 3D CNN–LSTM, 3D Bi-LSTM–CNN, CNN–Dilated Convolution–LSTM, and a Hybrid Ensemble CNN-LSTM integrated with a Decision Tree classifier. The proposed models are designed to combine spatial feature extraction, bidirectional sequence learning, multi-scale temporal modelling, and ensemble-based decision optimisation.

The novelty of this study lies in the unified comparison of these four architectures using the same micro-Doppler representations, data-processing stages, dataset partitioning, and evaluation metrics. The study also introduces the integration of dilated convolution and LSTM for long-range temporal learning, alongside an ensemble CNN-LSTM and Decision Tree strategy intended to improve classification stability and generalisation. Accordingly, this study aims to evaluate and compare the proposed architectures in terms of accuracy, precision, recall, F1-score, confusion matrix, inference latency, model parameters, and computational complexity. This research is important because practical HAR systems must operate without physical contact, preserve privacy, distinguish similar activities, and remain reliable under noisy and changing environmental conditions. The findings are expected to support the development of FMCW radar-based HAR systems for elderly monitoring, healthcare, rehabilitation, smart homes, security, and human–machine interaction.

METHODS

This study employed a comparative experimental design to evaluate four spatial–temporal deep-learning architectures for FMCW radar-based Human Activity Recognition: 3D CNN–LSTM, 3D Bi-LSTM–CNN, CNN–Dilated Convolution–LSTM, and a Hybrid Ensemble CNN-LSTM integrated with a Decision Tree classifier. Data were collected indoors using a 77 GHz FMCW radar while several participants performed seven activities: walking, running, sitting, standing, falling, picking up objects, and hand waving, with variations in body orientation and movement speed. The radar signals were processed through static-clutter removal, Range FFT, Doppler FFT, and Short-Time Fourier Transform to generate micro-Doppler spectrograms as model inputs, following the time–frequency processing approach described by Tan et al. (2024).

The spectrograms were resized, normalised, and enhanced using Hamming windowing, Gaussian filtering, and noise reduction. Data augmentation, including random cropping, time shifting, frequency masking, Gaussian noise injection, and temporal scaling, was applied to reduce overfitting. CNN and 3D CNN layers were used for spatial feature extraction, whereas LSTM and Bi-LSTM networks modelled temporal dependencies, consistent with the approach adopted by Wu et al. (2025). Dilated convolution was incorporated to enlarge the receptive field and capture multi-scale temporal patterns without a substantial increase in computational complexity, as proposed by Kakuba et al. (2024).

The dataset was divided into training, validation, and testing subsets at proportions of 70%, 15%, and 15%, respectively. All models were trained for 100 epochs using the Adam optimiser, a learning rate of 0.001, a batch size of 32, ReLU activation, a dropout rate of 0.5, and cross-entropy loss. Model performance was assessed using accuracy, precision, recall, F1-score, and confusion matrices, while computational efficiency was evaluated through parameter count, FLOPs, memory consumption, training time, and inference latency. These measures were used to identify the architecture offering the most appropriate balance between classification performance, generalisation, and computational efficiency.

RESULT AND DISCUSSION

This section presents a comprehensive quantitative and qualitative evaluation of the proposed hybrid spatial–temporal deep learning architectures for FMCW radar-based Human Ac-tivity Recognition (HAR). The experimental analysis compares the proposed models against conventional CNN architectures, standalone LSTM networks, and several existing radar HAR approaches reported in the literature. Performance evaluation focuses on spatial representation learning, temporal depen-dency modeling, long-sequence learning capability, general-ization performance, computational efficiency, and robustness under noisy environments.

Performance Comparison of Proposed Architectures

Table V summarizes the overall classification performance of all proposed architectures and baseline models. The results demonstrate that all proposed hybrid architec-tures significantly outperform conventional CNN and stan-dalone LSTM models. The Hybrid Ensemble CNN-LSTM combined with Decision Tree classification achieves the high-est recognition accuracy of 97.63%, indicating strong robust-ness and improved generalization capability Figure 14 illustrates the comparative accuracy performance of all evaluated models.

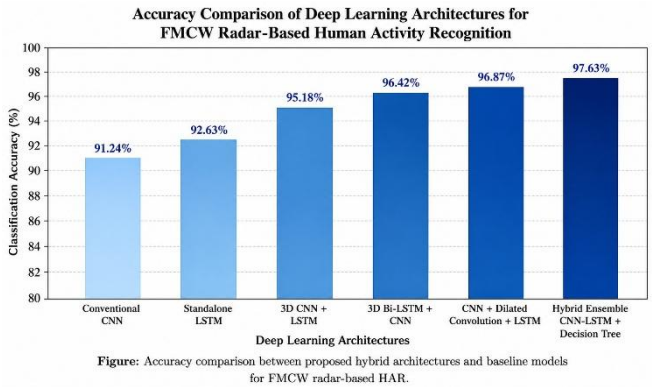

Figure 1. Accuracy comparison between proposed architectures and baseline models.

Comparison with Existing Radar HAR Approaches

The proposed frameworks were further compared with existing FMCW radar-based HAR approaches, including CNN–LSTM models for joint spatial–temporal feature learning (Qian et al., 2023), Bi-LSTM networks for continuous activity classification (Shrestha et al., 2020), asymmetric residual CNNs for micro-Doppler feature extraction (Zhang et al., 2024), and CNN–Bi-LSTM architectures for recognising activities with similar motion patterns (Wu et al., 2025). As presented in Table VI, the proposed Hybrid Ensemble CNN-LSTM–Decision Tree framework outperformed the compared methods. This improvement can be attributed to its integrated optimisation of spatial feature extraction, temporal-dependency modelling, receptive-field expansion, and ensemble-based decision-making. The combination of complementary spatial and temporal representations also reduces prediction variance and improves classification stability, consistent with the benefits of ensemble learning described by Rokach (2010).

Spatial Learning Effectiveness

Spatial feature learning is essential for extracting discriminative motion patterns from micro-Doppler spectrograms. Conventional CNNs effectively capture local spatial characteristics through

hierarchical convolutional operations, enabling the identification of activity-specific radar signatures (Zhang et al., 2024). However, spatial-only models remain limited in representing the complex temporal transitions inherent in sequential human movements, which motivates the integration of recurrent and convolutional components (Qian et al., 2023; Wu et al., 2025). In the proposed framework, 3D CNN operations are applied across both spatial and temporal dimensions, allowing the models to learn richer motion representations and preserve dynamic information within successive radar spectrograms. Figure 15 presents visualization examples of learned spatial feature maps extracted from different convolutional layers.

Table 1. Performance Comparison of Proposed Architectures and Baseline Models

Model	Accuracy	Precision	Recall	F1-Score	Inference Latency
Conventional CNN	91.24%	90.87%	90.12%	90.48%	12 ms
Standalone LSTM	92.63%	92.11%	91.75%	91.92%	18 ms
3D CNN + LSTM	95.18%	94.92%	94.71%	94.81%	24 ms
3D Bi-LSTM + CNN	96.42%	96.11%	95.98%	96.04%	29 ms
CNN + Dilated Conv + LSTM	96.87%	96.53%	96.44%	96.48%	26 ms
Hybrid Ensemble CNN-LSTM + DT	97.63%	97.32%	97.11%	97.21%	31 ms

Table 2. Comparison with Existing FMCW Radar-Based HAR Methods

Method	Architecture	Dataset	Accuracy
Previous Work A	CNN	FMCW Radar	91.8%
Previous Work B	CNN-LSTM	Micro-Doppler	94.5%
Previous Work C	Conv-LSTM	Radar HAR	95.2%
Previous Work D	Bi-LSTM	Radar Sequence	95.7%
Proposed Framework	Hybrid Ensemble CNN-LSTM + DT	FMCW Radar	97.63%

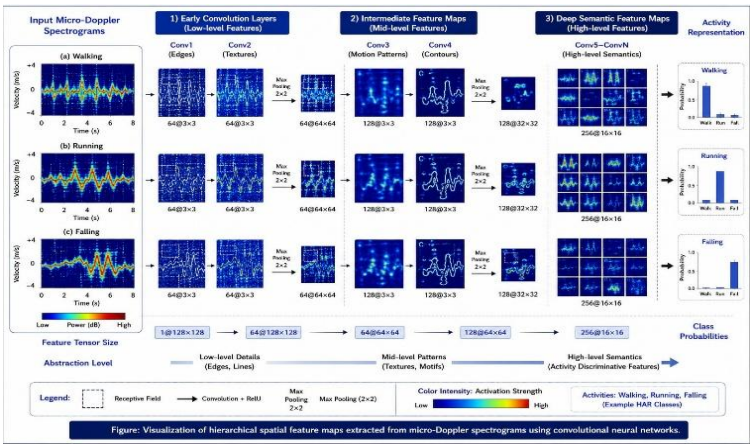

Figure 2. Visualization of spatial feature maps extracted from micro-Doppler spectrograms.

The extracted feature maps demonstrate that deeper convo-lutional layers successfully capture higher-level motion repre-sentations corresponding to various human activities.

Temporal Dependency Learning

Temporal dependency modelling is essential in radar-based HAR because human activities consist of successive motion stages rather than isolated spatial patterns. Conventional CNNs can extract local features from micro-Doppler spectrograms but cannot adequately preserve long-term relationships between consecutive radar frames. LSTM addresses this limitation through input, output, and forget gates that regulate the retention of relevant historical information (Hochreiter & Schmidhuber, 1997). Bi-LSTM extends this capability by processing sequences in both forward and backward directions, enabling the model to interpret each motion pattern using its preceding and subsequent contexts.

The experimental results showed that the proposed 3D Bi-LSTM–CNN achieved an accuracy of 96.42%, precision of 96.11%, recall of 95.98%, and F1-score of 96.04%, with an inference latency of 29 ms. Its accuracy exceeded that of the conventional CNN and standalone LSTM by 5.18 and 3.79 percentage points, respectively. This improvement indicates that combining three-dimensional spatial–temporal feature extraction with bidirectional sequence modelling provides a more complete representation of activity transitions. Although the hybrid ensemble model obtained the highest overall accuracy, the 3D Bi-LSTM–CNN demonstrated a particularly strong capacity for temporal representation because it simultaneously considered past and future motion states.

These findings are consistent with Shrestha et al. (2020), who reported that a Bi-LSTM applied to continuous Doppler-domain radar sequences achieved a mean accuracy above 90%, whereas range-domain data achieved approximately 76%. Qian et al. (2023) similarly combined parallel LSTM networks with three-dimensional convolution to integrate temporal and spatial information from multiple radar spectrograms, demonstrating the benefit of joint feature learning. More recently, Wu et al. (2025) reported that a CNN-BiLSTM model achieved 94.63% accuracy when distinguishing four similar human activities, confirming that bidirectional temporal information is valuable when motion signatures overlap. Direct numerical comparisons should nevertheless be interpreted cautiously because the studies employed different datasets, activity classes, participants, radar configurations, and validation protocols. Figure 16 illustrates the temporal dependency learning mechanism of the proposed Bi-LSTM architecture.

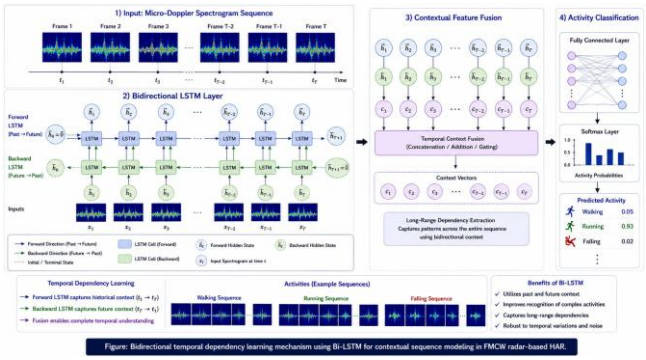

Figure 3. Bidirectional temporal dependency learning using Bi-LSTM archi-tecture.

The results indicate that Bi-LSTM significantly improves contextual representation because the network can utilize information from both past and future temporal states simulta-neously. This bidirectional learning mechanism enables more accurate recognition of activities involving complex temporal transitions.

Long-Sequence Modeling Capability

Long-sequence dependency extraction remains one of the major challenges in radar HAR systems. Standard convolu-tional operations possess limited receptive fields, restricting the capability of CNN architectures to capture large-scale temporal structures. The proposed CNN + Dilated Convolution + LSTM frame-work effectively addresses this limitation through receptive field expansion using dilated convolution operations. Different dilation rates enable the network to capture broader temporal contexts without significantly increasing computational com-plexity. Figure 3 illustrates receptive field expansion achieved using dilated convolution.

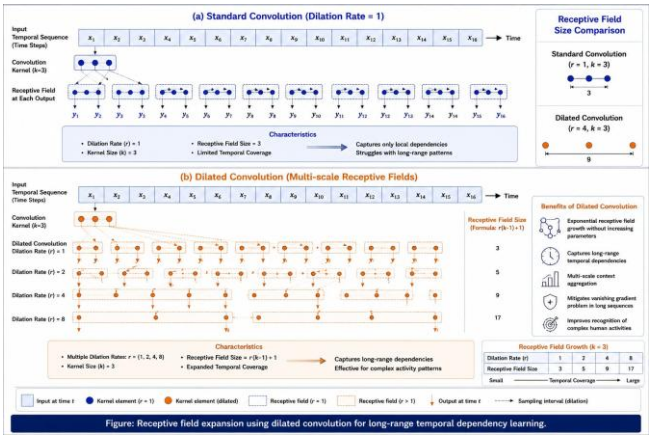

Figure 3. Receptive field expansion using dilated convolution for long-range temporal dependency learning.

Dilated convolution improves temporal awareness by en-abling the model to learn multi-scale temporal representations while preserving computational efficiency. Furthermore, di-lated convolution mitigates vanishing gradient problems by al-lowing more direct long-range information propagation across deep temporal layers. Experimental results demonstrate that the CNN + Dilated Convolution + LSTM framework achieves superior long-sequence modeling capability compared to conventional CNN-LSTM architectures.

Generalization Capability and Overfitting Mitigation

Generalisation capability is a fundamental requirement for practical Human Activity Recognition systems operating across diverse participants and changing environmental conditions. Deep architectures with large parameter counts are particularly vulnerable to overfitting when trained on limited radar datasets; therefore, the proposed frameworks employ dropout regularisation, data augmentation, batch normalisation, spectrogram normalisation, and ensemble learning to improve robustness. Among the evaluated models, the Hybrid Ensemble CNN-LSTM–Decision Tree framework demonstrated the strongest generalisation performance by combining complementary spatial and temporal feature representations while reducing prediction variance and sensitivity to individual model errors, consistent with the advantages of ensemble classifiers described by Rokach (2010). The relatively small difference between training and validation performance further indicates that the adopted regularisation strategies effectively controlled overfitting and enabled the model to maintain stable recognition performance on previously unseen data.

Computational Tradeoffs

Although hybrid spatial–temporal architectures deliver substantial gains in classification accuracy, robustness, and temporal representation, these improvements are accompanied by higher computational demands than those of standalone CNN models. The inclusion of recurrent layers, bidirectional processing, dilated convolution, feature-fusion mechanisms, and multiple classifiers increases the number of trainable parameters, floating-point operations, memory usage, and inference time. As summarised in Table 2, this trade-off is most evident in the Hybrid Ensemble CNN-LSTM–Decision Tree framework, which achieves the strongest recognition performance but also requires the greatest computational resources and produces the highest inference latency among the evaluated architectures.

Nevertheless, the additional computational cost is justified by the framework's superior ability to integrate complementary spatial and temporal representations, reduce prediction variance, and maintain stable performance under noisy and variable sensing conditions. In practical intelligent-sensing applications, model selection should therefore not be based solely on computational efficiency, but on the required balance between accuracy, latency, robustness, and deployment constraints. Lightweight models may remain preferable for edge devices with strict memory and real-time processing limitations, whereas the Hybrid Ensemble framework is more suitable for safety-critical applications in which reliable activity recognition and generalisation are more

important than minimal computational overhead. Overall, the findings confirm that the proposed ensemble architecture offers the most effective performance complexity balance when high recognition reliability is the primary system requirement.

Discussion

The results demonstrate that integrating spatial and temporal learning mechanisms produces consistently stronger FMCW radar-based Human Activity Recognition performance than using either mechanism independently. The Hybrid Ensemble CNN–LSTM–Decision Tree achieved the highest accuracy of 97.63%, together with 97.32% precision, 97.11% recall, and a 97.21% F1-score. Compared with the conventional CNN, the ensemble framework improved accuracy by 6.39 percentage points, while its improvement over the standalone LSTM was 5.00 percentage points. It also exceeded the 3D CNN–LSTM, 3D Bi-LSTM–CNN, and CNN–Dilated Convolution–LSTM models by 2.45, 1.21, and 0.76 percentage points, respectively. These results indicate that the performance gain was not produced by a single component, but by the complementary contribution of spatial feature extraction, sequential modelling, receptive-field expansion, and ensemble-based decision optimisation.

The findings are broadly consistent with previous research emphasising the importance of combining spatial and temporal representations. Qian et al. (2023) employed parallel LSTM networks and three-dimensional convolution to learn temporal and spatial information from multiple radar spectrograms, demonstrating that joint feature extraction is more effective than relying on a single representation. Shrestha et al. (2020) similarly showed that LSTM and Bi-LSTM networks can process continuous micro-Doppler sequences and retain information across activity transitions. More recently, Wu et al. (2025) combined CNN and Bi-LSTM to distinguish activities with similar radar signatures, whereas Zhang et al. (2024) strengthened spatial feature extraction through asymmetric convolutional residual blocks. The present study extends these approaches by evaluating multiple spatial–temporal configurations under the same processing pipeline and by introducing ensemble decision learning as an additional classification stage.

Within the comparison presented in this study, the proposed ensemble framework outperformed the selected CNN, CNN–LSTM, Conv-LSTM, and Bi-LSTM reference models. Nevertheless, comparisons with results reported across different studies must be interpreted with care. Tan et al. (2024), for example, reported an accuracy of 98.2% using a two-stream 1D-CNN, cross-channel operation, and BiGRU architecture on the Rad-HAR dataset. This value is slightly higher than the 97.63% obtained in the present study, but the two results are not directly equivalent because the datasets, activity classes, radar configurations, preprocessing procedures, and validation protocols differ. Accordingly, the principal contribution of the present research lies not in claiming universal superiority over every published model, but in demonstrating the relative effectiveness of four advanced architectures under a consistent experimental design.

The comparison between the conventional CNN and 3D CNN–LSTM provides evidence that spatial features alone are insufficient to represent complete human-motion dynamics. The conventional CNN achieved 91.24% accuracy, whereas the 3D CNN–LSTM reached 95.18%, representing an improvement of 3.94 percentage points. Conventional two-dimensional convolution identifies local structures within individual spectrograms, but it does not explicitly preserve how these structures change across successive radar frames. By applying convolution across spatial and temporal dimensions, the 3D CNN captures short-term motion evolution before the resulting features are processed by the LSTM. This mechanism explains why the deeper feature maps produced more distinctive representations of the seven activity classes. The result supports Zhang et al. (2024), who demonstrated the importance of strengthening spatial micro-Doppler feature extraction, while also confirming the conclusion of Qian et al. (2023) that spatial representations become more informative when combined with sequential learning.

Temporal modelling produced a further improvement in activity-transition recognition. The 3D Bi-LSTM–CNN achieved 96.42% accuracy and a 96.04% F1-score, outperforming both the standalone LSTM and 3D CNN–LSTM. Its advantage is attributable to bidirectional processing, through which each radar segment is interpreted using information from preceding and subsequent motion states. This is particularly relevant for activities such as sitting, standing, falling,

and picking up an object, as their early motion patterns may overlap before diverging in later stages. Shrestha et al. (2020) demonstrated that Bi-LSTM networks are effective for continuous FMCW radar sequences because they preserve temporal relationships rather than interpreting radar data solely as independent images. Wu et al. (2025) also found that CNN–Bi-LSTM modelling improved the discrimination of activities with similar motion signatures. The present findings reinforce these observations and show that bidirectional temporal modelling becomes more effective when it is combined with three-dimensional feature extraction.

A particularly important result was obtained by the CNN–Dilated Convolution–LSTM model, which achieved 96.87% accuracy with an inference latency of 26 ms. This model exceeded the 3D Bi-LSTM–CNN by 0.45 percentage points while reducing latency by 3 ms. The result suggests that expanding the receptive field can provide a computationally efficient alternative for capturing long-range temporal structures. Dilated convolution allows non-adjacent temporal features to be processed without proportionally increasing kernel size or network depth, thereby preserving information across broader movement intervals. Kakuba et al. (2024) similarly applied dilated causal convolution to radar point-cloud data to capture temporal relationships across extended sequences. In the present framework, its combination with CNN and LSTM enabled the model to learn local spectrogram characteristics, multi-scale temporal contexts, and sequential dependencies within a unified architecture.

The superior performance of the Hybrid Ensemble CNN–LSTM–Decision Tree further indicates that different learning modules captured complementary aspects of human motion. CNN components emphasised local spectrogram structures, LSTM components preserved sequential information, and the Decision Tree provided an additional decision boundary for the fused feature representation. Ensemble learning can reduce dependence on errors produced by an individual model and improve prediction stability by integrating multiple representations, as explained by Rokach (2010). The close correspondence between the ensemble model's precision, recall, and F1-score also indicates internally consistent classification performance rather than an improvement restricted to only one evaluation metric. Furthermore, the reported small difference between training and validation performance suggests that dropout, batch normalisation, spectrogram normalisation, data augmentation, and ensemble learning collectively limited overfitting.

The improvement in recognition performance was accompanied by greater computational demand. The conventional CNN required only 12 ms per inference, whereas the Hybrid Ensemble framework required 31 ms. The additional latency resulted from recurrent processing, feature fusion, and ensemble classification. However, the increase of 19 ms over the conventional CNN yielded a 6.39-percentage-point improvement in accuracy and substantially stronger precision, recall, and F1-score. This trade-off is acceptable for applications in which reliable recognition is more important than minimum latency, including fall detection, elderly monitoring, rehabilitation, and safety surveillance. For highly constrained edge devices, the CNN–Dilated Convolution–LSTM may offer a more balanced alternative because it achieved 96.87% accuracy with lower latency than the ensemble and Bi-LSTM models.

The computational results also highlight that model selection should reflect the intended deployment environment. Lim et al. (2024) achieved 95.54% accuracy using a lightweight radar point-cloud model containing only 22.28 thousand parameters and demonstrated that quantisation could substantially reduce memory consumption for edge implementation. Their findings show that practical HAR deployment requires consideration of accuracy, memory, power use, and hardware availability rather than accuracy alone. In this context, the proposed Hybrid Ensemble framework is most appropriate for systems with adequate computational resources or safety-critical requirements, whereas future compressed versions could employ quantisation, pruning, or knowledge distillation for low-power deployment.

The study offers both methodological and practical implications. Methodologically, the results show that radar HAR performance benefits from a hierarchical learning strategy in which local spatial patterns, short-term motion changes, long-range dependencies, and classifier-level decisions are optimised jointly. The strong result of the dilated-convolution model also suggests that temporal receptive-field design should receive the same attention as recurrent depth when

developing radar HAR architectures. Practically, the 97.63% ensemble accuracy and 31 ms inference latency indicate its potential for responsive non-contact sensing systems. Radar-based monitoring is particularly relevant in privacy-sensitive environments because it identifies motion signatures without recording identifiable visual images, as demonstrated in daily-activity monitoring research by Diraco et al. (2025).

Several boundaries of the present evaluation also define opportunities for further validation. The experiments were conducted in a controlled indoor setting using seven activity classes and a fixed 77 GHz FMCW radar configuration. Future studies could strengthen external validity by including more participants, additional daily activities, multiple rooms, varied sensor positions, and subject-independent testing protocols. Cross-dataset evaluation would also determine whether the learned representations remain stable when radar hardware and acquisition conditions change. Although data augmentation and regularisation supported generalisation, dedicated experiments involving interference, severe multipath conditions, and unseen environments would provide a more detailed measurement of noise robustness. These extensions would complement, rather than diminish, the current findings, which already provide a consistent comparison of four hybrid architectures under identical experimental conditions.

Finally, the study evaluated inference latency and computational complexity at the model level, but deployment-specific energy consumption was not measured on embedded hardware. Future implementation on edge platforms would enable direct assessment of power consumption, memory use, and sustained real-time performance. Repeated training with confidence intervals, statistical significance testing, and component-wise ablation analysis could also quantify the individual contributions of 3D convolution, Bi-LSTM, dilated convolution, augmentation, and ensemble classification. Overall, the findings provide strong evidence that the Hybrid Ensemble CNN–LSTM–Decision Tree offers the most reliable overall classification performance, while the CNN–Dilated Convolution–LSTM provides a competitive alternative when a lower-latency performance–complexity balance is required.

CONCLUSION

This study developed and comparatively evaluated four hybrid spatial–temporal deep-learning architectures for FMCW radar-based Human Activity Recognition using micro-Doppler spectrograms. The results confirm that combining spatial feature extraction with temporal dependency modelling produces more reliable recognition than conventional CNN and standalone LSTM models. Among the evaluated architectures, the Hybrid Ensemble CNN-LSTM–Decision Tree achieved the strongest overall performance, with an accuracy of 97.63%, precision of 97.32%, recall of 97.11%, and F1-score of 97.21%. The 3D CNN components effectively captured discriminative spatial–temporal motion patterns, Bi-LSTM improved bidirectional contextual modelling, and dilated convolution expanded the receptive field for multi-scale temporal learning. Although the ensemble framework required greater computational resources and an inference time of 31 ms, its improved classification stability and generalisation demonstrate an appropriate trade-off for applications in which recognition reliability is the primary requirement. Overall, the proposed framework provides a robust and privacy-preserving alternative for intelligent healthcare monitoring, elderly assistance, smart homes, non-visual surveillance, and human–machine interaction.

Future research should validate the proposed architectures using larger and more diverse participant groups, subject-independent protocols, multiple environments, additional activity classes, and multi-person scenarios. Transformer and attention-based mechanisms may be investigated to strengthen global temporal modelling and improve the interpretability of micro-Doppler representations, while multimodal fusion with thermal, Wi-Fi, LiDAR, or inertial sensors may enhance robustness in complex environments. For practical deployment, practitioners should select the Hybrid Ensemble framework when safety and recognition reliability are prioritised, while the CNN–Dilated Convolution–LSTM architecture may be more suitable for resource-constrained edge devices. Model compression, quantisation, pruning, and knowledge distillation should also be considered to reduce memory use and inference latency. At the policy level, the adoption of radar-

based HAR should be supported by standards governing performance validation, data security, privacy protection, interoperability, and responsible deployment, particularly in healthcare and public-monitoring environments.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to the Faculty of Electrical Engineering, Telkom University, for providing academic support and research facilities throughout this study. The authors also thank all participants involved in the FMCW radar data-acquisition process and colleagues who contributed valuable technical feedback during the development and evaluation of the proposed Human Activity Recognition framework.

REFERENCES

Chen, V. C. (2019). *The micro-Doppler effect in radar* (2nd ed.). Artech House.

Chen, Z., Sun, Y., & Qu, L. (2025). Research on cross-scene human activity recognition based on radar and Wi-Fi multimodal fusion. *Electronics, 14*(8), 1518. https://doi.org/10.3390/electronics14081518

Cohen, L. (1995). *Time-frequency analysis*. Prentice Hall.

Diraco, G., Rescio, G., & Leone, A. (2025). Radar-based activity recognition in strictly privacy-sensitive settings through deep feature learning. *Biomimetics, 10*(4), 243. https://doi.org/10.3390/biomimetics10040243

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16 × 16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kakuba, S., Colaco, S. J., Kim, J. H., Lee, D.-G., Yoon, Y., & Han, D. S. (2024). Dilated causal convolution based human activity recognition using voxelized point cloud radar data. In *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 812–815). IEEE. https://doi.org/10.1109/ICAIIC60209.2024.10463502

Kim, Y., & Ling, H. (2009). Human activity classification based on micro-Doppler signatures using a support vector machine. *IEEE Transactions on Geoscience and Remote Sensing, 47*(5), 1328–1337. https://doi.org/10.1109/TGRS.2009.2012849

Lara, O. D., & Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials, 15*(3), 1192–1209. https://doi.org/10.1109/SURV.2012.110112.00192

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lim, S., Park, C., Lee, S., & Jung, Y. (2024). Human activity recognition based on point clouds from millimeter-wave radar. *Applied Sciences, 14*(22), 10764. https://doi.org/10.3390/app142210764

Lin, Y., Li, H., & Faccio, D. (2024). Human multi-activities classification using mmWave radar: Feature fusion in time-domain and PCANet. *Sensors, 24*(16), 5450. https://doi.org/10.3390/s24165450

Qian, Y., Chen, C., Tang, L., Jia, Y., & Cui, G. (2023). Parallel LSTM-CNN network with radar multispectrogram for human activity recognition. *IEEE Sensors Journal, 23*(2), 1308–1317. https://doi.org/10.1109/JSEN.2022.3224083

Qian, Y., Chen, C., Tang, L., Jia, Y., & Cui, G. (2023). Parallel LSTM-CNN network with radar multispectrogram for human activity recognition. *IEEE Sensors Journal, 23*(2), 1308–1317. https://doi.org/10.1109/JSEN.2022.3224083

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33*(1–2), 1–39. https://doi.org/10.1007/s10462-009-9124-7

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 28, pp. 802–810). Curran Associates.

Shrestha, A., Li, H., Le Kernec, J., & Fioranelli, F. (2020). Continuous human activity classification from FMCW radar with Bi-LSTM networks. *IEEE Sensors Journal, 20*(22), 13607–13619. https://doi.org/10.1109/JSEN.2020.3006386

Skolnik, M. I. (2001). *Introduction to radar systems* (3rd ed.). McGraw-Hill.

Tan, T.-H., Tian, J.-H., Sharma, A. K., Liu, S.-H., & Huang, Y.-F. (2024). Human activity recognition based on deep learning and micro-Doppler radar data. *Sensors, 24*(8), 2530. https://doi.org/10.3390/s24082530

Trinh, T. H., Dai, A. M., Luong, M.-T., & Le, Q. V. (2018). Learning longer-term dependencies in RNNs with auxiliary losses. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 4965–4974). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v80/trinh18a.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 5998–6008). Curran Associates.

Wu, X., Ling, Z., Zhang, X., Ma, Z., & Deng, W. (2025). Human similar activity recognition using millimeter-wave radar based on CNN-BiLSTM and class activation mapping. *Eng, 6*(3), 44. https://doi.org/10.3390/eng6030044

Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*. https://arxiv.org/abs/1511.07122

Zhang, Y., Tang, H., Wu, Y., Wang, B., & Yang, D. (2024). FMCW radar human action recognition based on asymmetric convolutional residual blocks. *Sensors, 24*(14), 4570. https://doi.org/10.3390/s24144570

Zhao, M., Adib, F., & Katabi, D. (2016). Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking* (pp. 95–108). Association for Computing Machinery. https://doi.org/10.1145/2973750.2973762