

Article

Victim Detection and Localization for Search-and-Rescue: A Robot-Mounted UWB Radar with a Hybrid CNN–ViT Model

Antonios-Periklis Michalopoulos ^{1,*}, Efstratios N. Paliodimos ², Grigoris Nikolaou ¹, Demetrios Cantzos ¹ and Stylianos A. Mytilinaios ²

¹ Department of Industrial Design and Production Engineering, University of West Attica, 12244 Aigaleo, Greece; nikolaou@uniwa.gr (G.N.); cantzos@uniwa.gr (D.C.)

² Department of Electrical and Electronics Engineering, University of West Attica, 12244 Aigaleo, Greece; mscres-86@uniwa.gr (E.N.P.); smitil@uniwa.gr (S.A.M.)

* Correspondence: amichalopoulos@uniwa.gr

Abstract

Robotic systems for search-and-rescue operations require robust, non-line-of-sight victim detection in order to locate trapped individuals behind obstacles with high precision. This paper presents a robotic victim-localization system based on a convolutional neural network—vision transformer (CNN–ViT) architecture trained on an open-source radar dataset for through-wall presence detection. In addition to binary presence detection, the proposed approach uses attention information from the transformer layers to estimate the region of the radar signal most relevant to the victim location. The model is deployed on a robotic platform and tested in an additional realistic environment, where classification and distance-estimation outputs are fused into a heatmap-style spatial representation. This enables the system to localize the estimated victim position on the map generated by the robot. To enhance robustness, the system is evaluated using both a leave-one-subject-out (LOSO) protocol on the original open-source radar dataset and additional experimental sessions collected with the robotic platform. On the original dataset, the model achieved victim-detection F1 scores of 82–96% and distance-estimation MAE of 0.25–0.65 m relative to the robot. On newly collected, previously unseen data, it achieved F1 scores of 72–92% and an MAE of 0.16–0.78 m on correctly classified present samples. The complete pipeline was further deployed on a mobile robot in an environment different from the one used to collect the original dataset, illustrating the potential of the proposed system for practical search-and-rescue scenarios.

Keywords: search-and-rescue; vision transformers; robotics; victim localization



Academic Editor: Feng Wei

Received: 7 June 2026

Revised: 17 July 2026

Accepted: 22 July 2026

Published: 24 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Search-and-Rescue (SAR) operations are constantly evolving as new hardware and software tools are introduced to help teams locate survivors faster and operate more safely. A challenge in real deployments is finding people who are not in direct line-of-sight, e.g., behind walls, inside partially collapsed rooms, or under debris. Because time is critical, physically breaching structures and manually inspecting occluded areas can be slow, risky, and often impractical during the early phase of response. For these reasons, radar sensing and mobile robots have been widely explored in SAR research and practice. Through-wall radars can detect human presence, in most cases based on the respiration-related micro-motions without requiring direct visibility [1], while robotic

platforms can enter hazardous or confined spaces that are unsafe for rescuers and provide remote situational awareness [2,3]. The synergy of these two systems, i.e., a radar-equipped ground robot, offers a way to scan obstructed areas more efficiently, while keeping operators at a safer distance.

In the literature, Ultra-Wideband (UWB) radar is mainly used because its wide bandwidth enables fine range discrimination while achieving practical penetration in indoor structures. A representative set of studies by the same research group [4–6] demonstrates how UWB impulse radar can support SAR-relevant “signs-of-life” sensing by exploiting respiration-induced micro-motions when there is no direct line of sight. In these works, human presence and distance are inferred from time–frequency models and respiration estimation is attained through denoising/decomposition models combined with frequency-accumulation estimation, enabling detection across different wall materials and clutter conditions. Other studies have combined UWB radars with machine learning to detect or localize people in non-line-of-sight conditions, such as behind walls [7] or through vegetation/foilage [8], using convolutional neural networks (CNN). Related work also exists for FMCW radar, where CNN-based methods have been applied to through-wall human sensing and detection/localization tasks [9,10]. However, many of these approaches are typically evaluated in controlled experimental settings and are not validated as complete SAR-ready systems (e.g., robot integration, deployment constraints, and testing across diverse real-world obstructions).

Beyond standalone radar experiments, only a limited number of studies have investigated robot-mounted radar systems to support through-wall sensing. Currently, to the best of our knowledge, apart from our previous work [11,12] no prior studies have addressed this problem through the combination of a robotic system and a radar, employing a multi-angle person search which results in a more robust localization of the subject as multiple sessions help to improve the position estimation. Studies closely related to our previous and current work, although they do not employ machine learning (ML) or deep learning (DL), include three systems. The work of [13] mounts a UWB radar on a drone to collect respiration signatures of subjects under debris in a laboratory environment. The aim of that work is to address an aerial search, buried-victim scenario rather than through-wall sensing and reports respiration signatures without utilizing a detection or localization model. The work of [14] combines a Bluetooth-controlled wheeled robot with a Doppler radar module and a PIR sensor using simple rule-based logic, and therefore provides only binary human-presence detection, without range or position estimation. The most complete study of these [15] integrates an SFCW radar into a semi-autonomous rescue robot and reports a full signal processing pipeline for multi-person detection, 2D-MUSIC localization, and breathing rate estimation, together with a public benchmark dataset and code; its localization, however, relies on a classical array-based (2D-MUSIC) method rather than a learned model, and cross-subject or out-of-distribution generalization is not evaluated. In summary, even though these systems demonstrate the use of radar-equipped robots for human sensing, none couples active multi-view acquisition with a learned model that jointly performs detection, distance estimation, and map-level victim localization as in our work.

Our previous [11,12] and current work aims to mitigate this gap in the literature by mounting the radar on a robotic platform to collect breathing signatures of lying victims from different angles and by utilizing DL models to estimate the subject’s location behind the wall. Although CNNs and transformer-based architectures have already been applied to radar-based human sensing, e.g., human activity recognition from distributed radar using combined vision transformer and CNN models [16], activity recognition from range-distributed time–Doppler sparse point clouds using a multi-channel Point-

Net [17], and DETR-based multi-view indoor perception from horizontal and vertical radar heatmaps [18], these works do not address robot-mounted UWB through-wall victim localization from breathing-related signatures. Therefore, the novelty of the proposed approach lies in combining CNN-based feature extraction with transformer attention for UWB through-wall human detection and attention-based range estimation. The main objective of the model is to improve generalization across different ranges, including distances longer than those used during training, for both classification and distance estimation. In addition, the proposed model is integrated with a 2D localization methodology that projects the estimated victim position onto the robot's map, enabling interpretation within the live mapping representation of the surrounding environment.

The main contributions of this work are summarized as follows:

- Hybrid CNN–ViT model: CNN-based feature extraction is combined with transformer attention for through-wall human detection, together with attention-based range estimation.
- Generalization across ranges: The model is designed and evaluated to generalize to distances longer than those seen in training, for both presence classification and distance estimation (LOSO and previously unseen data).
- Map-level victim localization: A 2D localization method projects the estimated victim position onto the robot's live map, allowing interpretation within the surrounding environment. The work was first proposed in [11] but is now presented more analytically and tested in a different lab environment.

2. Materials and Methods

2.1. Hardware

As a continuation of our previous works [11,12], we used the same experimental setup to ensure reproducibility; therefore, the robotic platform and radar sensor, the two primary components of this system, remain unchanged. The radar used in our experiments is the SLMX4 module from Sensor Logic [19], which is based on Novelda's X4 chip and operates at a center frequency of 7.29 GHz with a 1.4 GHz bandwidth. The antenna provides wide-angle coverage, with 3 dB beamwidths of 87.5° in elevation and 114.3° in azimuth. A more detailed description of the radar parameters is provided in [11,12].

The robotic platform integrates multiple sensing modalities to support accurate motion control, right-wall following, and consistent data acquisition from different observation angles for subject detection. In addition to the SLMX4 radar, the system incorporates a Microsoft Kinect V1 RGB-D camera, which is primarily used for navigation and can also provide visual feedback to a human operator when manual intervention is required. The Kinect V1 captures RGB-D data at up to 30 frames per second with a spatial resolution of 640×480 pixels and a depth range of up to 4.5 m. An LDS-01 LiDAR sensor [20] is further employed for wall-following control, allowing the robot to maintain a constant distance from the wall and improve the repeatability of radar measurements.

2.2. Data

The dataset consists of radar measurements acquired for through-wall human presence and breathing-related detection. Two data sources were used. First, an open-source radar-breathing dataset from our previous work [12] was employed for model training, validation, and baseline evaluation. Second, additional radar recordings were collected using the same experimental configuration to assess the generalization ability of the proposed models under conditions not fully represented in the original dataset.

The main objective of the data collection strategy was to maintain consistency with the original acquisition protocol while extending the evaluation scenarios. In particular,

the additional measurements were designed to include different subject distances and orientations, allowing the models to be tested beyond the range and viewing conditions used during training. This approach enables both controlled model development and a more realistic assessment of robustness in out-of-distribution scenarios.

2.2.1. Dataset

We use the open-source radar-breathing dataset from [12] to train and evaluate the models in our work. The dataset was collected with a robotic radar-sensing platform in the Antenna Laboratory of the EEE Department at the University of West Attica (UNIWA), using six participants (ages 22–26) with varied physical characteristics. Measurements were acquired across 11 robot positions along a wall, with the participants positioned in 4 orientations, with the robot placed 30 cm from a 27 cm thick YTONG wall while a subject lay at 1 m and 2 m behind the wall facing it; the robot was repositioned every 20 cm using an odometer to track displacement, emulating a search-and-rescue inspection trajectory. At each position, the complex radar backscatter envelope was recorded as time-stamped 2D matrices (frames of size 1×180 range bins), totalling 300,960 frames per subject and approximately 26.4 h of presence data across all subjects (≈ 19 FPS), along with an additional 5.5 h of absence (no-subject) recordings collected under the same configuration.

To further evaluate our models, we collected 33 different sessions shown in Table 1, with the same parameters as the dataset we used, at different distances and orientations. These data were used to evaluate the model performance in out-of-scope results, to assess how robustly the proposed models generalize to larger distances and different viewing orientations beyond the original dataset. The data collected included 9 absence sessions and 24 presence sessions. Most presence measurements were concentrated at short range, with 9 out of 24 sessions falling between 1 and 3 m, which corresponds to the range mainly represented in the training data. At the same time, 15 out of 24 presence sessions were located outside this short-range band, i.e., above 3 m. These out-of-range cases were mainly distributed across the mid-range distances, with 6 out of 24 sessions between 3 and 5 m and 7 out of 24 sessions between 5 and 7 m. Only a small number of sessions corresponded to longer distances, with 2 out of 24 sessions between 7 and 9 m. The overall ground-truth distances ranged from 1.63 m to 9 m.

Table 1. Test sessions collected for extrapolation to longer ranges and new viewpoints.

Range (Meters)	Count
1–3 m	9
3–5 m	6
5–7 m	7
7–9 m	2
Absent	9
Total	33

2.2.2. Data Manipulation

As in previous studies using the selected dataset, part of the dataset contains wall-induced noise. Following the description in prior work, we removed 28 range bins in the present study. The removal of even more bins occurred due to the aim of the current work to extend the range and further remove the noise in the first meters. So, from here on, the input data for the models will start from 1.433 m; this corresponds to 152 retained range bins.

The raw radar recordings were first transformed into structured range–time representations suitable for both classification and localization. Each measurement frame contains

complex-valued radar samples over successive range bins; therefore, the complex backscatter signal was converted into its amplitude representation in order to describe the strength of the received reflections at each time instant and range position. Since static reflections from the wall, surrounding objects, and hardware-related offsets can introduce a strong stationary component, DC removal was applied independently to each range bin. To preserve time-varying components, such as respiration-induced micro-motions, while suppressing constant or slowly invariant background contributions, for a radar amplitude signal $x_r(t)$ at range bin r , the new signal is computed as:

$$x'_r(t) = x_r(t) - \frac{1}{N} \sum_{t=1}^N x_r(t) \quad (1)$$

where N is the number of temporal samples in the recording. After this step, the data were temporally and spatially sliced to retain the range and time intervals relevant to the through-wall detection task.

To further emphasize breathing-related dynamics, the DC-corrected signals were filtered along the temporal axis using a Butterworth bandpass filter. The selected passband was designed to retain low-frequency physiological motion associated with human respiration while attenuating very slow baseline variations and higher-frequency noise. In this study, the filtering stage used a third-order Butterworth bandpass filter with cut-off frequencies of 0.07 Hz and 0.4 Hz, applied independently to each range bin. The normalized cut-off frequencies were obtained with respect to the Nyquist frequency as,

$$f_{c,n} = \frac{f_c}{f_s} f_c \in \left\{ f_{low}, f_{high} \right\} \quad (2)$$

where f_s is the radar sampling frequency and f_c denotes either the lower or upper cut-off frequency of the Butterworth bandpass filter.

Finally, to address the class imbalance inherent in the radar dataset, a comprehensive augmentation strategy, specifically targeting the absence class, was implemented. The absence samples, representing scenarios where no target was detected, were originally underrepresented compared to presence samples across all subject-configuration combinations. We applied four complementary augmentation techniques to artificially expand the absence dataset while preserving its statistical characteristics: (1) range shifting by randomly rolling the range dimension by ± 20 bins to simulate range-position variability in clutter/noise, (2) range scaling by multiplying signals by a random factor between 0.8 and 1.2 to represent different reflectivity properties, (3) white noise injection uniformly distributed across a $\pm 20\%$ range of the signal's mean value to simulate sensor noise, and (4) pink noise generation using $1/f$ filtering to capture natural clutter and environmental variations. Each original absence sample was processed to generate one augmented counterpart, effectively doubling the absence data while maintaining the original unaugmented version, resulting in a balanced dataset where the augmented samples are explicitly marked to enable controlled training and validation strategies.

To evaluate the individual contribution of each preprocessing step, an ablation study was conducted. Table 2 outlines the performance progression from the “No range-bin removal” to the proposed full pipeline, excluding filtered results, which are detailed in Section 3. Each ablation stage was trained using the model architecture proposed in Section 2.3 under the 70-20-10 split. Each stage isolated a specific preprocessing component; for instance, the “No DC removal” stage still utilized both range-bin removal and absence augmentation, as indicated by the respective range-bin counts and sample sizes.

Table 2. Testing ablation on preprocessing choices.

Stage	Range Bins	Test Accuracy	Test Precision	Test Recall	Test F1	Distance RMSE (m)	Distance MAE (m)	Number of Instances
No range-bin removal	180	0.7948	0.8480	0.8517	0.8499	1.2804	1.0372	1009
No DC removal	152	0.7740	0.7995	0.8924	0.8434	1.1738	1.0001	1009
No absence augmentation	152	0.8859	0.9350	0.9201	0.9275	1.0387	0.4369	868
Full pipeline	152	0.8860	0.9195	0.9128	0.9161	0.7603	0.3060	1009

Preliminary results indicate that both range-bin removal and DC removal play a pivotal role in optimizing both classification accuracy and distance estimation. The performance drop observed without range-bin removal is expected, as early range bins capture static environmental clutter (such as wall reflections) which introduces misleading features. On the other hand, at first the absence augmentation seemingly shows better results, but the number of instances is different than the other stages. Specifically, the other stages, and specifically the full pipeline, handle a significantly bigger sample of data, 141 additional instances, while maintaining almost the same classification accuracy. When factoring in its significantly more accurate distance estimation, the full pipeline emerges as the more robust and effective pipeline.

To prevent information leakage among the training, validation, and test sets, the data are partitioned at the recording-session level rather than at the individual-sample level, with each session identified by a unique code. All samples originating from the same session, including their augmented variants, are assigned to the same subset according to this code, so no augmented copy of a sample can appear in a different split from its original. For example, if session 104 is placed in the validation set, all of its original and augmented samples are assigned to validation as well. In the LOSO setting, an entire subject is held out for testing, which by construction prevents both temporal and augmented leakage, since no window from the held-out subject appears in training. In the non-LOSO setting, the same session-level rule is applied to the 70-20-10 split, ensuring that temporally adjacent windows from a single continuous recording cannot be distributed across different subsets.

2.3. Methods

ViTs have increasingly been explored as alternatives to CNNs in many different fields [21–23] and have shown promise for radar data analysis [24] due to their superior ability to capture global dependencies across sparse, high-dimensional signals [25] such as micro-Doppler maps. Unlike CNNs, which rely on local receptive fields, ViTs employ self-attention mechanisms to model relationships between all patches of the input simultaneously, enhancing tasks such as target recognition, semantic segmentation, and object detection. This global-context awareness is particularly useful for radar data, where relevant patterns may be distributed across time and range. In our case, relative position bias allows the model to encode spatial relationships between patches, which may support more robust range-aware representations when training data are available only at predefined distances. However, pure ViT architectures may struggle to efficiently capture fine-grained local structures, such as weak reflections, small Doppler variations, and compact scattering patterns, which are often critical in radar signals. Since the input is divided into patches before attention is applied, important local details may be degraded, especially when training data is limited. This limitation motivates the use of hybrid CNN–ViT architectures, where CNN layers first extract robust local features and transformer blocks subsequently model global dependencies across the radar representation.

For the localization of the victim, we experimented with several approaches as in our previous works [11,12]. Specifically, we tested several methods presented in our previous

work, some based on machine-learning models such as XGBoost, and one mathematical approach applied after the creation of features from the raw data. The results for the mathematical approach, with the use of the Standard Deviation (STD) on the data, achieved an MAE of 0.38 m and 0.49 m in both works and, in one work, with the use of XGBoost, 0.08 m. All the results can be seen in Section 3, alongside the results of the proposed solution.

Although XGBoost attained the lowest error, this result was obtained on a subset of the data spanning a narrow range of distances, and it can only be trusted under such conditions: as a tree-based regressor, it cannot predict ranges outside the interval covered by its training labels. The STD approach, by contrast, requires no retraining and generalizes across the full range of distances, which makes it a stronger conventional baseline. However, it is sensitive to clutter and noisy environments, where the variation induced by the target's breathing can be masked or mimicked by interfering reflections, and it depends on a separate, hand-crafted processing pipeline decoupled from the detection model. The attention-rollout estimator avoids both drawbacks: it is obtained directly from the classifier already used for presence detection, within the same single forward pass and with no dedicated regression training.

2.3.1. Victim Detection

The proposed architecture is a hybrid convolutional neural network and vision transformer (CNN-ViT) model designed for person detection with an auxiliary attention-based range-estimation mechanism, derived from the ViT attention rollout, that does not require training a separate regression head. The model receives as input a two-dimensional range–time radar heatmap of shape $(470 \times 152 \times 1)$, where the first axis corresponds to 470 consecutive time steps at a sampling rate of 18.7 Hz (approximately 25 s of observation), and the second axis corresponds to 152 range bins spanning a physical distance of 1.433 m to 9.216 m at a range resolution of 0.0512 m per bin. The single channel contains the amplitude of the complex IQ radar signal after DC removal.

Prior to any learned transformation, the input range–time radar heatmap is subjected to a Layer Normalization operation computed independently per sample over all spatial (height and width) and channel dimensions (H, W, C). Without this normalization, the model may over-rely on absolute amplitude differences caused by range, clutter, sensor gain, or environmental conditions rather than learning breathing-related temporal patterns.

The normalized input is processed by a three-stage convolutional backbone. Each stage consists of a convolutional block comprising two consecutive Conv2D layers with 3×3 kernels and the same padding, each followed by Batch Normalization and ReLU activation. Following the double convolution, each block applies a Squeeze–Excitation (SE) mechanism. The SE block performs a global average pooling over the spatial dimensions to produce a channel descriptor vector, which is passed through two fully connected layers. The three convolutional blocks are configured with filter counts of 16, 32, and 64 respectively. The first two blocks apply a 2×2 max pooling operation following the SE mechanism, progressively halving the spatial dimensions. The third block deliberately omits pooling to preserve spatial resolution for the subsequent FPN-inspired multi-scale fusion. Spatial dropout with rates of 0.05, 0.10, and 0.10 is applied after pooling in each block respectively, zeroing entire feature maps rather than individual values to encourage redundancy and prevent co-adaptation across channels.

Rather than discarding intermediate CNN representations and passing only the deepest feature map to the subsequent transformer encoder, the fusion module fuses all three convolutional stages into a single unified feature map. The motivation is that each CNN layer encodes qualitatively different information: shallow layers retain fine-grained spatial detail, while deeper ones capture more abstract semantic representations of the features.

The fusion is performed as follows. Each intermediate representation is projected to a common channel dimensionality of 64 via a 1×1 convolution:

$$p_3 = W_3^{1 \times 1} * c_3, \quad p_2 = W_2^{1 \times 1} * c_2, \quad p_1 = W_1^{1 \times 1} * \text{AvgPool}(c_1) \quad (3)$$

where c_1 is spatially downsampled via 2×2 average pooling to match the spatial dimensions of c_2 and c_3 at 117×38 . The three projected feature maps are then fused by element-wise addition:

$$f = p_1 + p_2 + p_3 \quad (4)$$

followed by a final 1×1 convolution, Batch Normalization, and ReLU activation to produce a coherent fused representation of $117 \times 38 \times 64$.

The fused feature map is then cut into a grid of non-overlapping 3×2 tiles, which results in 741 patches. Each patch consists of $3 \times 2 \times 64$ values. Each of the 741 flattened vectors (384 numbers) is linearly projected onto 64 dimensions via a shared weight matrix. Then a learnable classification token (CLS token), initialized to zero and updated during training, is prepended to the patch sequence. The resulting sequence contains 742 tokens, 741 patches and one CLS token, with 64 values per token.

The token sequence of 742 tokens is then processed by four successive transformer encoder blocks. Each block applies a Pre-Layer Normalization followed by a Multi-Head Self-Attention (MHSA) with 4 heads and a head dimension of 16. Within each block, every token is projected onto three vectors (Q, K, V) through a single shared linear layer. The attention score between any two tokens is computed as the dot product of their query and key vectors, with a learnable relative position bias added to encode the spatial offset between patches in the 39×19 grid. These scores are passed through a softmax to produce attention weights, and each token's output is a weighted sum of all value vectors across the sequence. The outputs of all 4 heads are concatenated and passed through a final linear projection. Each block additionally applies a two-layer MLP with a hidden dimension of 128 and GELU activations independently to each token, followed by residual connections around both the attention and MLP sub-layers. Stochastic depth regularization is applied to each residual branch with a linearly increasing drop probability from 0.0 at the first block to 0.2 at the final block. The attention weights produced by each block are cached internally after every forward pass for use in the distance-estimation procedure described later.

Following the final transformer block, the token sequence is passed through a Layer Normalization to produce the final representation. The first representation is the CLS token, taken directly from position 0 of the sequence, which, after attending to all 741 patch tokens across all 4 transformer layers, now carries a focused discriminative summary of the entire scene. The second is the global average of all 741 patch tokens, which provides a spatially distributed summary of the full spectrogram. The two vectors, each of 64 dimensions, are concatenated into a single 128-dimensional representation. This combined vector is passed through a two-layer MLP with hidden dimensions of 64 and 32, GELU activations, and a dropout rate of 0.2 applied after each layer. A final dense layer with a sigmoid activation produces a single scalar output between 0 and 1, where values above 0.5 indicate person presence and values below 0.5 indicate an empty room. The model is trained using binary cross-entropy loss with class-balanced sample weighting to compensate for imbalances between presence and absence samples. Figure 1 is a visual representation of the described model architecture.

The model was trained using the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4} , combined with a binary cross-entropy loss function. Training was monitored using multiple evaluation metrics including binary accuracy, area under the curve (AUC), precision, and recall. To prevent overfitting and ensure training stability, we

implemented early stopping with a patience of 12 epochs, monitoring validation loss, and a learning rate reduction scheduler (ReduceLROnPlateau) with a patience of 10 epochs.

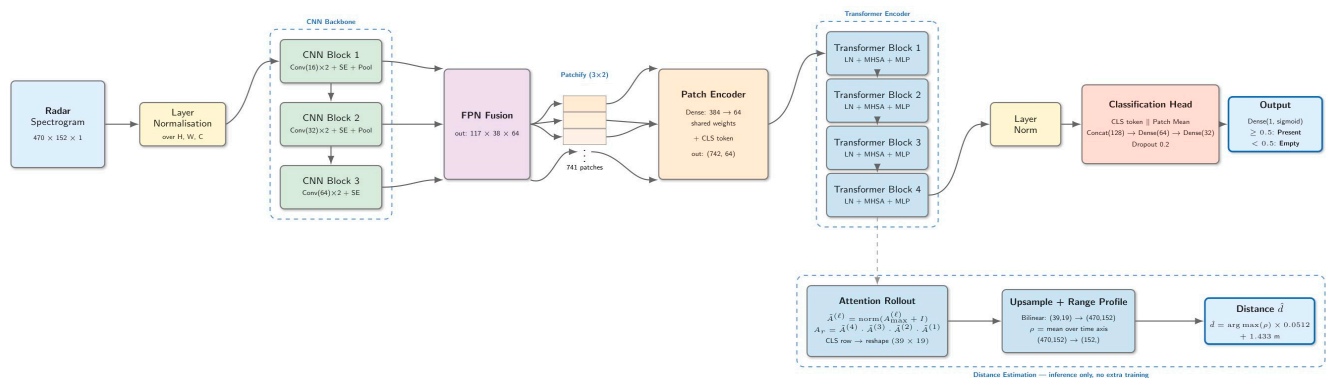


Figure 1. Visual representation of the proposed CNN-ViT architecture.

2.3.2. Victim Localization

The attention weights produced by each of the four transformer blocks are cached internally after every forward pass, yielding four attention tensors, one per block. The four cached attention tensors are combined into a single matrix through the attention-rollout procedure. For each block, the attention tensor is reduced across the four heads by taking the element-wise maximum, a residual identity term is added to account for the skip connections, and the result is row-normalized. The element-wise maximum is used to preserve the sharpest localized attention response across heads, since averaging can dilute weak breathing-related peaks, particularly at longer ranges. The four resulting matrices are then multiplied sequentially to produce a single joint matrix that encodes the cumulative attention flow through the entire depth of the transformer.

The CLS token is then extracted and yields a vector of 741 values, which is reshaped to a grid of 39×19 and is bilinearly upsampled to the original range–time resolution of 470×152 . For the final distance estimation, the upsampled map is collapsed along the time axis by taking, for each range bin, the maximum attention across all time frames, producing a one-dimensional range attention profile, whose peak bin is converted to a physical distance as

$$d = \operatorname{argmax}(p) \times 0.0512 + 1.433 \quad (5)$$

where 0.0512 m is the radar range resolution and 1.433 m corresponds to the physical distance of the first retained range bin after range cropping.

To determine the optimal method for selecting the maximum attention value across all time frames for distance estimation, a series of experiments was conducted to evaluate the generalization capability of different aggregation approaches. The test set and external datasets were evaluated using four different statistical methods for distance estimation: mean, standard deviation (std), median, and maximum (max). The results are presented in the following table.

The comparison includes all available experimental configurations obtained from the evaluated models, considering both leave-one-subject-out (LOSO) and non-LOSO settings, as well as filtered and non-filtered data. Based on the results obtained, the maximum attention aggregation method achieved the most consistent performance across the evaluated datasets. Therefore, the max-based approach was selected as the final method and was used for the subsequent experiments, as it achieved the smallest gap between internal and external error and the lowest external RMSE. All the results can be seen in Table 3 below.

Table 3. Comparison of attention value across all time frames for distance estimation for test set and external data.

Metric	Test Set RMSE (m)	External Data RMSE (m)
Median	0.9521	2.1132
Mean	1.0892	1.4952
Max	1.2860	1.0895
Std	1.3019	1.0916

2.3.3. Probabilistic Localization Heatmap Fusion

For accurate localization of detected subjects, we created a probabilistic localization heatmap that accumulates data across successive radar sessions as the robot moves alongside the right wall. The heatmap fusion is independent of the model architecture, requiring only presence confidence and a distance estimation. This allows the method to be integrated into any perception pipeline that provides these outputs for visualization in RViz. At each robot position, the perception pipeline outputs (i) a presence confidence from the classifier c and (ii) a range estimate from the attention-based range estimator $\hat{d}$. A heatmap is then instantiated in the global ROS map frame and published as an OccupancyGrid for RViz visualization visualized as a heatmap of the environment. The heatmap created at each sensing session is a directional circular sector (wedge) on the right side of the robotic system with ± 45 degrees from the center of the right side of the robotic system. For each sensing session, the wedge is centered on the robot's right-facing axis, and each grid cell is evaluated according to its Euclidean distance from the radar position and its angular position relative to this axis; cells outside the angular wedge are assigned zero intensity. Within the wedge, each cell is assigned an intensity according to its distance from the estimated subject range $\hat{d}$. By setting r as the distance of a grid cell and as $c \in [0, 1]$ the classifier confidence, the intensity I of each grid cell i is:

$$I_i(r) = c * \max(0, 1 - \frac{|r_i - \hat{d}|}{\hat{d}}) \quad (6)$$

Thus, the intensity is maximal at $r = \hat{d}$, having the highest intensity, decreasing linearly as $|r - \hat{d}|$ increases, becoming zero at $r = 0$ and $r = 2\hat{d}$. The intensity of the grid is color-coded with blue as the lowest value and mapped nonlinearly from blue to red for the higher values until 95%, where the color changes to cyan for better visualization.

To localize the subject reliably, multiple sensing sessions are required; as the robot moves and collects additional data, the corresponding wedges partially overlap in the global grid. The values of these overlapping cells must therefore be updated to reflect the combination of different sessions, where each session contributes a different intensity pattern. That can be achieved by using the standard probabilistic OR (union) rule. By having the H as the global heatmap and the S as the current heatmap for a different session t , we have the normalized values of them as:

$$h_{t-1}(i) = \frac{H_{t-1}(i)}{100}, s_t(i) = \frac{S_t(i)}{100} \quad (7)$$

where $h_{t-1}(i), s_t(i) \in [0, 1]$. The overlapping cells fused values can be computed using the standard probabilistic OR (union) rule:

$$h_t(i) = 1 - (1 - h_{t-1}(i))(1 - s_t(i)) \quad (8)$$

Finally, the final global heatmap is mapped back on the occupancy scale based on (7) and (8):

$$H_t(i) = \text{clip}(100h_t(i), 0, 100) \quad (9)$$

Figure 2 illustrates the simulated implementation of the progressive localization capabilities through heatmap evolution. Panels A1 and A2 depict empty space with no detections, yielding null heatmaps due to absence, as their values are below 50%, thus classified as absence. Panel B1 introduces a weak detection at 4 m with 60% intensity, signifying initial tentative contact from a single session lacking confirmatory overlap. B2 demonstrates enhanced confidence via overlapping areas, concentrating probability in shared areas, while C1 and C2 exhibit stronger detections with vivid red peaks in overlapping regions indicative of heightened reliability. Finally, D1 and D2 converge on precise victim localization through integrated high-confidence heatmaps, resolving prior ambiguities. All variables used for the simulation of Figure 2 are present in Table 4.

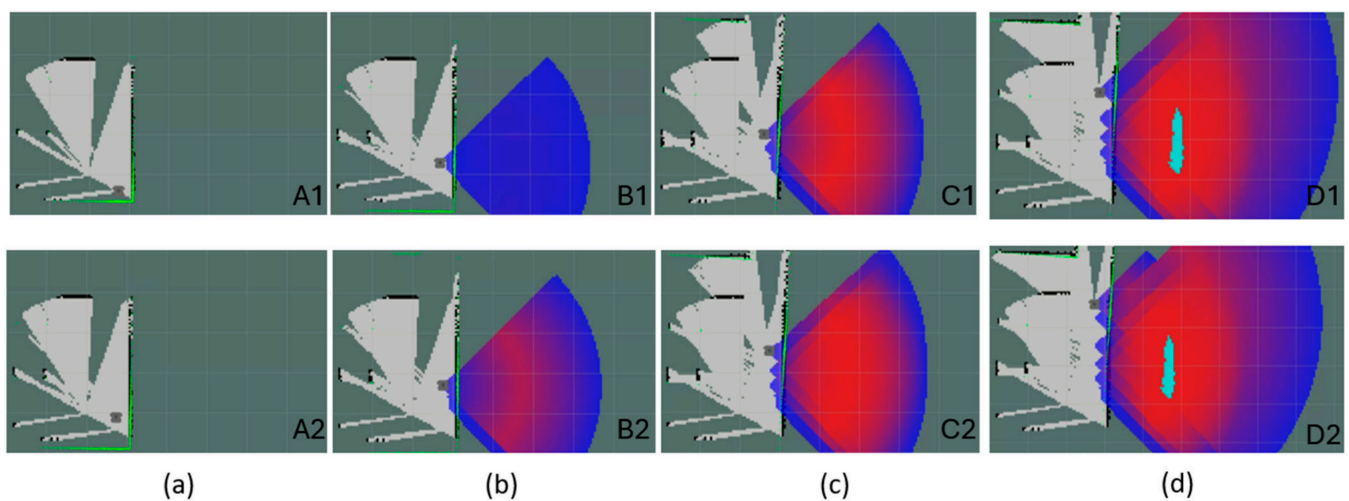


Figure 2. (a) Low-intensity examples with no visual heatmap; (b) first examples of medium-intensity with one overlap; (c) high-intensity examples due to overlapping instances; (d) final examples with overlapping intensity reaching over 95% (cyan).

Table 4. Simulation values used in the visualization of Figure 2.

Panel	Range (Meters)	Intensity (Percentage)
A1	-	10
A2	-	10
B1	4	60
B2	4	75
C1	4	90
C2	4	70
D1	6	75
D2	2	50

3. Results

The proposed models were evaluated using the original dataset and the additional sessions described earlier. These experiments were performed using either the data after applying the Butterworth filter described earlier or the minimally processed dataset, described in Section 2.2.2. For the presence detection task, we applied LOSO validation to assess model performance, while for range estimation, RMSE and MAE were computed only over presence samples that the model correctly classified as present; missed detections were

excluded, as they contribute to disproportionately large range errors. Absence samples are not included in the range-estimation metrics, since no ground-truth distance exists for them. Lastly, to validate the robustness of all approaches, we report results obtained from data outside the training range.

3.1. Dataset Results

The classification models that were trained and evaluated were trained under the LOSO protocol, with six LOSO models plus one non-LOSO model. Each model was also trained with and without Butterworth filtering, resulting in six unfiltered LOSO models, six filtered LOSO models, and two non-LOSO models, one using the filtered data and one using the unfiltered data.

The classification and range-estimation results, presented in Table 5, show that the proposed models achieved strong performance across both LOSO and non-LOSO evaluation settings. In the LOSO evaluation, the unfiltered models achieved an F1-score of 0.92 ± 0.05 and the filtered models 0.91 ± 0.02 , an essentially unchanged classification performance. The use of the Butterworth filter resulted in minimal change in LOSO classification performance, but it reduced the distance-estimation errors, with RMSE decreasing from 0.94 ± 0.35 m to 0.69 ± 0.27 m and MAE decreasing from 0.46 ± 0.10 m to 0.40 ± 0.13 m. In the non-LOSO evaluation, the filtered model outperformed the unfiltered model, increasing the F1-score from 0.92 to 0.94. It also reduced the distance errors, with RMSE decreasing from 0.76 m to 0.50 m and MAE from 0.30 m to 0.25 m.

Table 5. Model results on the test set are summarized.

Model Group	Accuracy	Precision	Recall	F1-Score	RMSE (m)	MAE (m)
LOSO-Unfiltered	0.90 ± 0.06	0.96 ± 0.02	0.88 ± 0.09	0.92 ± 0.05	0.94 ± 0.35	0.46 ± 0.10
LOSO-Filtered	0.89 ± 0.03	0.96 ± 0.01	0.88 ± 0.04	0.91 ± 0.02	0.69 ± 0.27	0.40 ± 0.13
Non-LOSO-Unfiltered	0.89	0.92	0.91	0.92	0.76	0.30
Non-LOSO-Filtered	0.91	0.94	0.94	0.94	0.50	0.25
2D-CNN + XGBoost [11] *	-	0.94	0.94	0.95	0.10	0.08
1D-CNN + STD [12]	-	0.78	0.80	0.77	0.88	0.49

* The results are only from a small sample of the dataset.

Overall, the results indicate that the Butterworth-filtered data mainly improved distance estimation and provided a clearer benefit in the non-LOSO setting, while LOSO classification performance remained relatively stable across the two preprocessing conditions.

For a further comparison of the proposed CNN-ViT model, we included results from previous works. Specifically, the 1D-CNN + STD model was evaluated using the same test set as the proposed model and achieved the lowest performance among the methods reported in Table 5. It should be noted that the reported results for the 1D-CNN + STD model correspond to the macro-average performance across all models, following the same evaluation approach used for the LOSO-Unfiltered and LOSO-Filtered experiments. The distance estimation with STD using the same model produced results that were not substantially different from the current results, while also representing a generalized method, since it can be applied directly to the data in its original mathematical form.

Lastly, the comparison with the 2D-CNN + XGBoost model is included only as a benchmark. It is not a fully representative comparison with the current work, because the previous model was evaluated on a much smaller portion of the dataset, limited to ranges below 2 m. The reason for including this model is to demonstrate that XGBoost can achieve good results within a known and limited data scope, but its performance should not be interpreted as evidence of generalization beyond that scope.

For the range-estimation metrics reported, we consider only the samples that were correctly classified by the model. That is, we measure how accurately the model localizes a person once it has correctly determined that a person is present. Restricting the analysis to correctly predicted samples decouples the range-estimation performance from classification errors, which would otherwise confound the two tasks. We acknowledge that reporting metrics on correctly predicted samples alone can be misleading, as it excludes the cases which the model failed to detect. For transparency, Table 6 additionally reports the same metrics computed over the entire dataset, allowing the reader to observe how much the results differ when all samples are included.

Table 6. Distance-estimation comparison with missed detections.

Fold	Detection		RMSE (m)			MAE (m)		
	Missed Detections	Missed Detections (%)	All Data	Correct Only	Reduction (%)	All Data	Correct Only	Reduction (%)
Unfiltered								
N1	62	5.5	0.89	0.80	10.4	0.46	0.43	6.3
N2	113	10.1	0.79	0.76	3.2	0.42	0.40	5.3
N3	56	5.1	0.76	0.73	4.3	0.51	0.48	4.5
N4	125	11.3	0.77	0.73	5.3	0.39	0.36	7.7
N5	100	9.0	1.07	0.99	8.0	0.47	0.41	13.8
N6	332	29.0	1.82	1.64	10.3	0.82	0.66	19.6
NO-LOSO	60	8.7	0.99	0.76	22.9	0.39	0.31	21.6
Filtered								
N1	77	6.8	0.59	0.58	1.7	0.31	0.30	0.8
N2	129	11.5	0.83	0.67	19.0	0.43	0.39	9.8
N3	122	11.0	0.80	0.66	17.3	0.46	0.43	6.0
N4	134	12.1	0.60	0.52	13.6	0.32	0.30	5.6
N5	150	13.6	0.82	0.50	39.5	0.42	0.35	17.8
N6	224	19.6	1.36	1.22	9.7	0.69	0.65	4.9
NO-LOSO	44	6.4	0.53	0.51	3.8	0.26	0.25	1.9

In Table 6, across every fold in both the unfiltered and filtered configurations, restricting the evaluation to correct predictions lowered both the RMSE and the MAE, confirming that misdetections contribute to disproportionately large range errors. The magnitude of this reduction was fold-dependent, ranging from below 2% to roughly 40% for RMSE. Notably, this reduction did not scale with the proportion of missed detections alone; the unfiltered NO-LOSO fold contained only 8.7% missed detections yet showed a 22.9% RMSE reduction, whereas the filtered N1 fold, with a comparable 6.8% of missed detections, showed only a 1.7% reduction.

A visualization from the test set is shown in Figure 3. The left panel presents the original radar image, while the right panel shows the corresponding attention overlay. The estimated distance is indicated by the dotted red line.

In this example, the original radar visualization shows a clear pattern around the 2.8 m range. The attention overlay indicates that the model assigns attention to regions associated with the breathing pattern at this distance. However, attention is also observed around 8 m, which is likely caused by environmental noise. This suggests that the model may also consider other recurring patterns in the radar data when determining whether breathing is present.

It should be noted that the attention overlay should not be interpreted as a deterministic explanation of the model's final classification decision. The model considers multiple features when producing its output. In this work, the attention overlay is used primarily to support the interpretation of the distance estimation rather than to provide a complete explanation of the classification result.

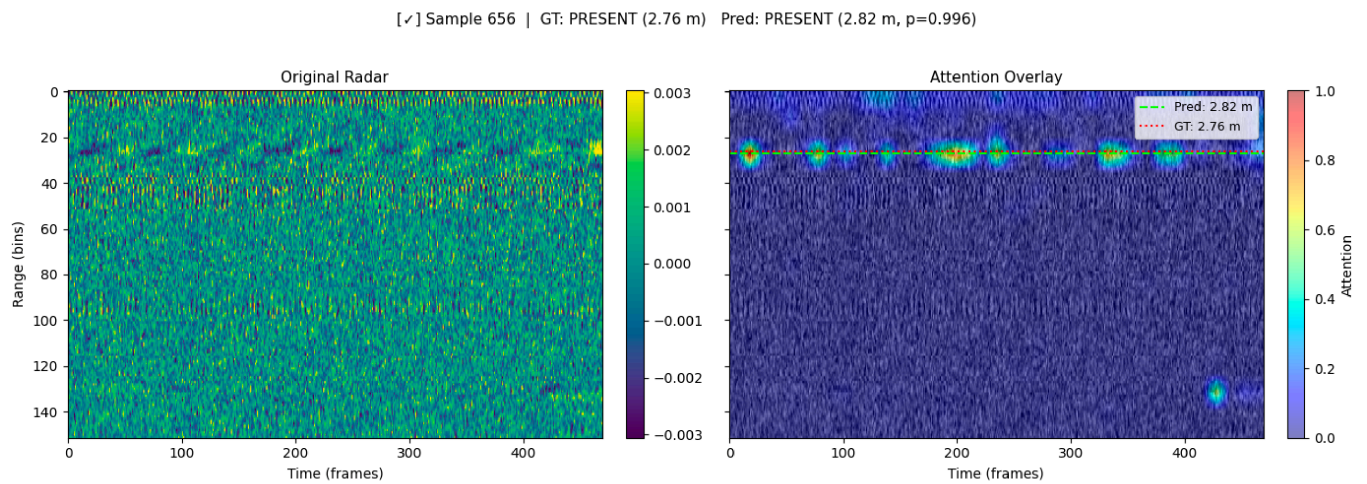


Figure 3. Example of the test set with breathing pattern at 2.82 m and noise at 8 m, with the correct prediction indicated by the ✓ symbol. The left panel shows the original radar signal used as input to the model, while the right panel presents the corresponding attention overlay together with the estimated distance. The model used in that example is the Non-LOSO-Unfiltered.

A more detailed analysis of the results in Table 7 showed that in the LOSO results, participant N1 achieved the best overall performance, with the highest test accuracy and F1-score in both the unfiltered and filtered conditions. In contrast, participant N6 consistently produced the weakest performance, with the lowest accuracy, recall, and F1-score. This indicates that N6 was the most difficult case for the model to generalize to, highlighting the effect of inter-participant variability in the LOSO setting, where each model is tested on a completely unseen participant.

Table 7. Model results on the test set for all the models.

Model Group	Fold/Split	Accuracy	Precision	Recall	F1-Score	RMSE (m)	MAE (m)
LOSO-Unfiltered	N1	0.9468	0.9736	0.9452	0.9592	0.7966	0.4306
LOSO-Unfiltered	N2	0.9164	0.9710	0.8989	0.9336	0.7605	0.4014
LOSO-Unfiltered	N3	0.9402	0.9571	0.9493	0.9532	0.7293	0.4835
LOSO-Unfiltered	N4	0.9069	0.9675	0.8870	0.9255	0.7294	0.3643
LOSO-Unfiltered	N5	0.8877	0.9161	0.9095	0.9128	0.9887	0.4087
LOSO-Unfiltered	N6	0.7895	0.9667	0.7098	0.8185	1.6361	0.6583
LOSO-Unfiltered	Mean ± SD	0.8979 ± 0.0574	0.9587 ± 0.0216	0.8833 ± 0.0886	0.9171 ± 0.0513	0.9401 ± 0.3545	0.4578 ± 0.1058
LOSO-Filtered	N1	0.9345	0.9679	0.9319	0.9495	0.5844	0.3032
LOSO-Filtered	N2	0.8994	0.9583	0.8846	0.9200	0.6746	0.3876
LOSO-Filtered	N3	0.8961	0.9452	0.8896	0.9166	0.6624	0.4344
LOSO-Filtered	N4	0.8981	0.9614	0.8788	0.9183	0.5179	0.3020
LOSO-Filtered	N5	0.8848	0.9531	0.8643	0.9065	0.4961	0.3468
LOSO-Filtered	N6	0.8480	0.9623	0.8042	0.8762	1.22	0.6518
LOSO-Filtered	Mean ± SD	0.8935 ± 0.0279	0.9580 ± 0.0079	0.8756 ± 0.0417	0.9145 ± 0.0237	0.6933 ± 0.2701	0.4043 ± 0.1315
Non-LOSO-Unfiltered	70/20/10	0.8860	0.9195	0.9128	0.9161	0.7603	0.3060
Non-LOSO-Filtered	70/20/10	0.9138	0.9374	0.9360	0.9367	0.5069	0.2526

Bold rows represent the summary statistics (mean ± standard deviation) calculated from the results of all LOSO folds (N1–N6).

3.2. Experiments

3.2.1. Distances Outside of the Scope

As mentioned earlier, our objective was to assess whether the model generalizes to ranges and conditions beyond those used during training. For this purpose, the trained models were applied to all 33 sessions of the external dataset described in Section 2.2.1, with the results summarized in Table 8 and reported per model in Table 9. These results indicate that the proposed methodology can detect human breathing patterns in data collected outside the N1–N6 dataset. Performance varied across configurations: the Non-LOSO-Unfiltered model retained internal-level performance, with an external F1 of 0.92 essentially

matching its internal value, whereas the remaining configurations showed lower and more variable F1 scores on the external data.

Table 8. Classification results on data collected and summarized.

Model Group	Accuracy	Precision	Recall	F1-Score	RMSE (m)	MAE (m)
LOSO-Unfiltered	0.83 ± 0.07	0.90 ± 0.02	0.87 ± 0.08	0.88 ± 0.05	0.61 ± 0.45	0.37 ± 0.22
LOSO-Filtered	0.78 ± 0.05	0.96 ± 0.02	0.74 ± 0.08	0.83 ± 0.05	0.25 ± 0.04	0.22 ± 0.04
Non-LOSO-Unfiltered	0.88	0.92	0.92	0.92	0.73	0.33
Non-LOSO-Filtered	0.67	0.93	0.58	0.72	0.20	0.16

Table 9. Classification results on data collected for all the models.

Model Group	Fold/Split	Accuracy	Precision	Recall	F1-Score	RMSE (m)	MAE (m)	Instances
LOSO-Unfiltered	N1	0.8788	0.9167	0.9167	0.9167	0.8550	0.4640	22
LOSO-Unfiltered	N2	0.8485	0.9130	0.8750	0.8936	0.3196	0.2573	21
LOSO-Unfiltered	N3	0.7273	0.8571	0.7500	0.8000	1.4313	0.7864	18
LOSO-Unfiltered	N4	0.8485	0.8800	0.9167	0.8980	0.3482	0.2587	22
LOSO-Unfiltered	N5	0.9091	0.9200	0.9583	0.9388	0.5021	0.3196	23
LOSO-Unfiltered	N6	0.7879	0.9048	0.7917	0.8444	0.2249	0.1719	19
LOSO-Unfiltered	Mean $\pm$ SD	0.83 ± 0.07	0.90 ± 0.02	0.87 ± 0.08	0.88 ± 0.05	0.61 ± 0.45	0.37 ± 0.22	-
LOSO-Filtered	N1	0.8182	0.9500	0.7917	0.8636	0.2124	0.1928	19
LOSO-Filtered	N2	0.8182	0.9500	0.7917	0.8636	0.2114	0.1952	19
LOSO-Filtered	N3	0.7576	0.9444	0.7083	0.8095	0.2846	0.2632	17
LOSO-Filtered	N4	0.8182	0.9500	0.7917	0.8636	0.2375	0.2167	19
LOSO-Filtered	N5	0.7879	0.9474	0.7500	0.8372	0.2116	0.1975	18
LOSO-Filtered	N6	0.6970	1.0000	0.5833	0.7368	0.3174	0.2975	14
LOSO-Filtered	Mean $\pm$ SD	0.78 ± 0.05	0.96 ± 0.02	0.74 ± 0.08	0.83 ± 0.05	0.25 ± 0.04	0.22 ± 0.04	-
Non-LOSO-Unfiltered	70/20/10	0.8788	0.9167	0.9167	0.9167	0.7333	0.3259	22
Non-LOSO-Filtered	70/20/10	0.6667	0.9333	0.5833	0.7179	0.1978	0.1633	14

Bold rows represent the summary statistics (mean $\pm$ standard deviation) calculated from the results of all LOSO folds (N1–N6).

However, these results should be interpreted with caution. A direct comparison with the full test set is limited, since the externally collected dataset contains substantially fewer samples than the internal test set. Nevertheless, the results provide preliminary evidence that the proposed approach can generalize beyond the original training distribution and can detect breathing activity in previously unseen data.

For further understanding of the results, a collection of three examples from the 33 sessions presented in Table 8 is presented below. All the examples use the Non-LOSO-Unfiltered model presented in Table 9, with the best F1-Score.

In the first example, shown in Figure 4, the breathing-related pattern appears at approximately 7.2 m. Although the breathing signature is not clearly visible in the range–time heatmap, a weak periodic pattern resembling respiration can be observed. The proposed model successfully assigned high confidence to this sample with human breathing activity, producing an 84% confidence score for person presence in the session. Moreover, the attention map focused on the same range region where the breathing-related pattern was located, indicating that the model based its prediction on the relevant part of the radar signal.

Figure 5 presents the second example from data outside the original dataset, where the breathing pattern is clearly visible at approximately 5.2 m. The left panel still shows the original radar signal used as input to the model, while the right panel shows the corresponding attention overlay and the estimated distance indicated by the dotted line. In this example, the model predicts the presence of a victim with high confidence (99.3%) and estimates the distance as 5.27 m. The attention overlay is concentrated around the breathing-related regions near the estimated distance. This suggests that the model can focus on temporal regions associated with breathing activity, although the attention map should be interpreted only as a qualitative visualization and not as a complete deterministic explanation of the model decision.

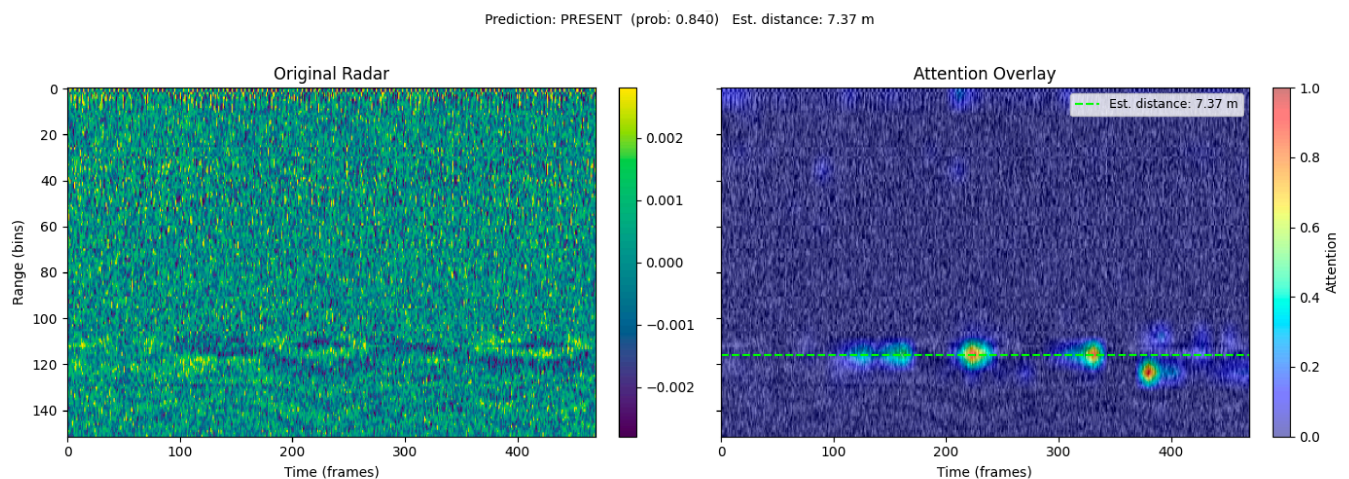


Figure 4. Example of data outside the scope of the original dataset with breathing pattern at 7.2 m. The left panel shows the original radar signal used as input to the model, while the right panel presents the corresponding attention overlay together with the estimated distance. The model used in that example is the Non-LOSO-Unfiltered.

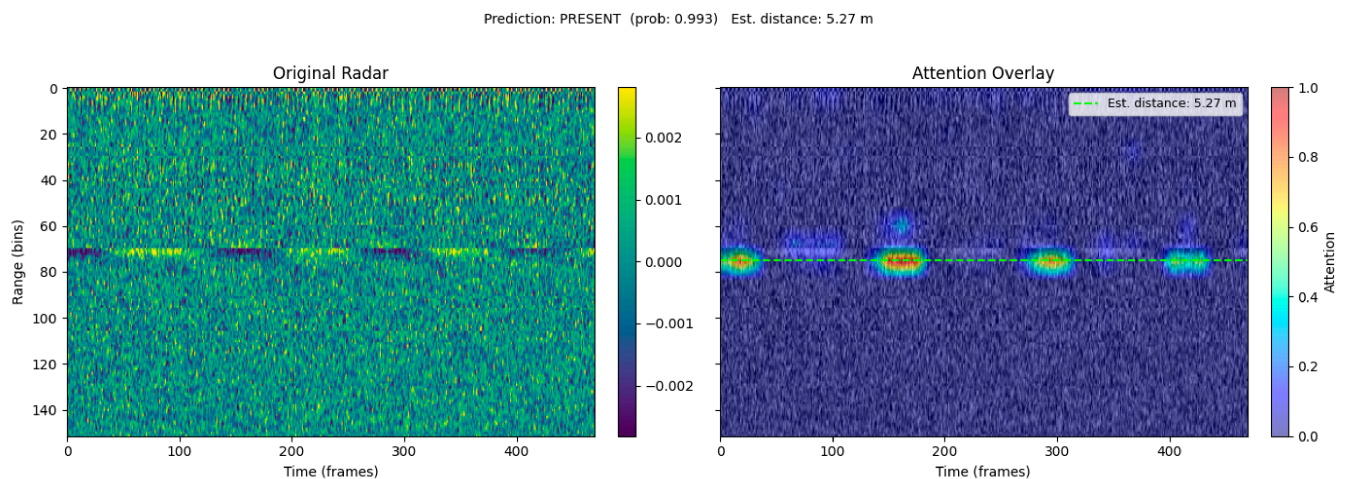


Figure 5. Example of data outside the scope of the original dataset with a breathing pattern at 5.2 m. The left panel shows the original radar signal used as input to the model, while the right panel presents the corresponding attention overlay together with the estimated distance. The model used in that example is the Non-LOSO-Unfiltered.

Lastly, Figure 6 presents an absence example, where no clear breathing pattern is visible in the original radar signal. Although the model assigns closer attention to a noisy region after approximately 400 frames, this region is not interpreted as breathing activity. As a result, the model correctly predicts the absence of a victim, with a low presence probability of 0.034 and no distance estimation, as the model classified that no person is present in the current sample. This example shows that high attention values do not necessarily correspond to a positive breathing detection, and that the model can distinguish between environmental noise and breathing-related patterns.

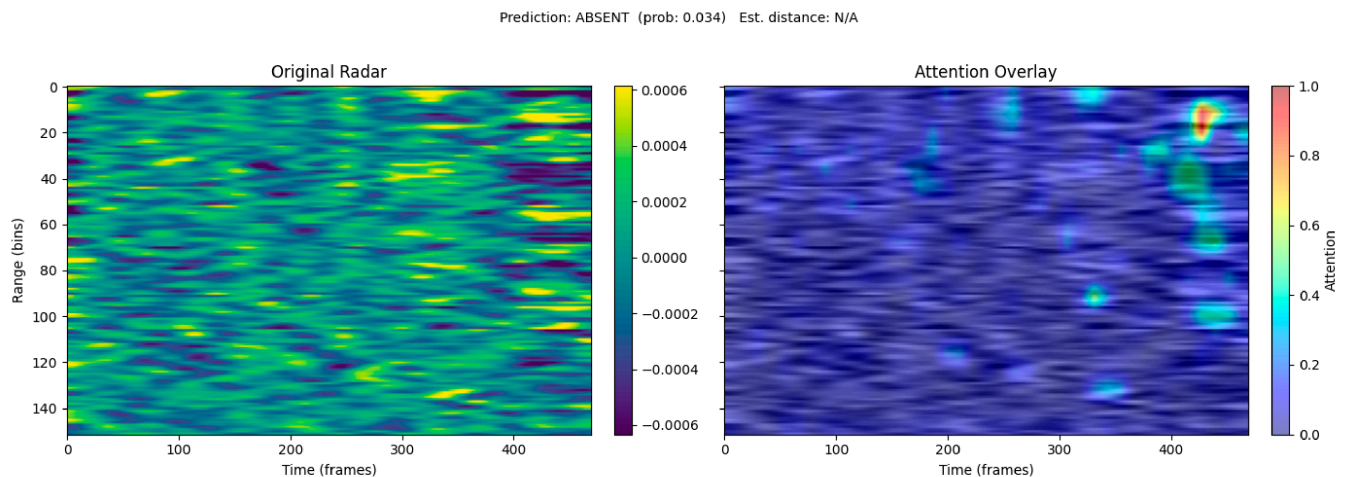


Figure 6. Example of data outside the scope of the original dataset with no breathing pattern present. The left panel shows the original radar signal used as input to the model, while the right panel presents the corresponding attention overlay. In the example here, there is no breathing pattern present, and so no distance estimation. The model used in that example is the Non-LOSO-Unfiltered.

3.2.2. Testing in Real Conditions

To evaluate the generalization capability of the proposed models, the trained models were deployed and tested in a previously unseen environment. The selected test environment was a different university space, distinct from the one used during training, characterized by a higher degree of clutter. The area contained various objects such as computers, whiteboards, and metallic chairs and tables, which introduced a more complex and noisier signal environment. In addition, several objects were positioned in close proximity to the walls, further attenuating the radar signal reflected from the target.

The robotic system navigated autonomously using a PID wall-following controller, tracking the right wall of the laboratory without any human intervention. At each designated position, the system remained static for 34 s while collecting radar data. The collected data were subsequently cropped to the appropriate input dimensions for the model before the robot proceeded to the next position.

The robot's starting position was on the left side of the laboratory exterior, approximately 2 m from the corner. As the robot progressed, it recorded data across four sessions along the initial corridor segment before reaching the turning point. During the turn, two additional sessions were recorded, during which the robot adjusted its lateral proximity to the right wall, resulting in two sessions captured at an oblique angle relative to the target. Following the turn, seven additional sessions were recorded until the conclusion of the experiment.

Figure 7 presents the results obtained using the Non-LOSO-Unfiltered model on the experimental data, while Figure 8 presents results from the same data using the Non-LOSO-Filtered model. The Non-LOSO-Unfiltered model (Figure 7) demonstrates more consistent detection of the subject with fewer spurious detections. In contrast, the Non-LOSO-Filtered model (Figure 8) exhibits a higher number of detections overall, with greater spatial dispersion in the heatmap, though the peak localization appears more precise. However, the Non-LOSO-Filtered model also introduces a greater number of outlier detections compared to the Non-LOSO-Unfiltered counterpart.

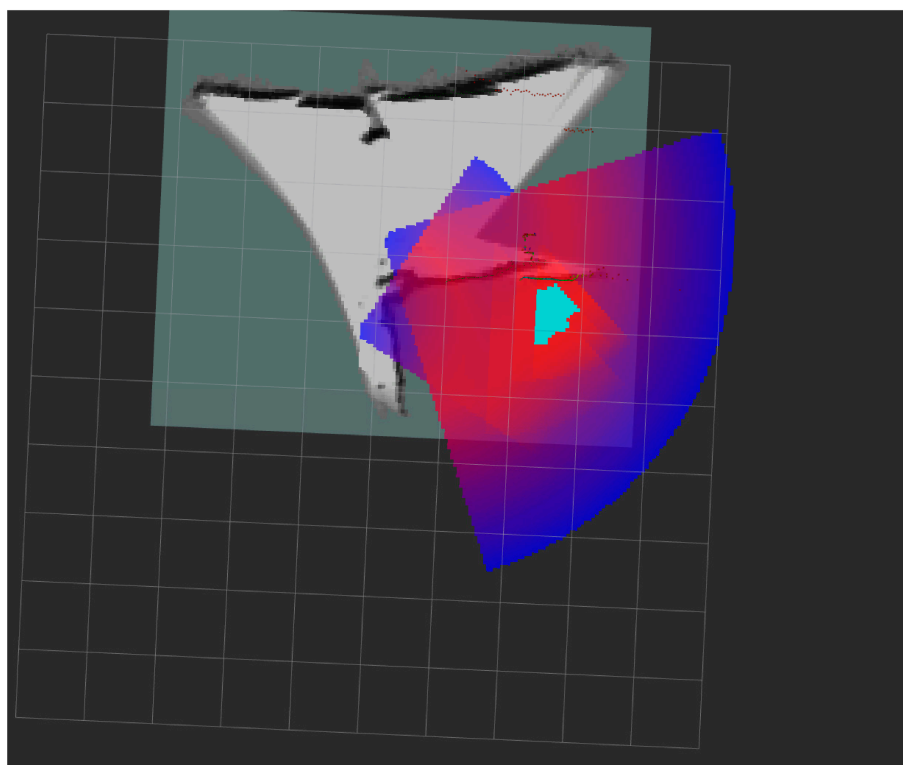


Figure 7. Results from the experiment. The model used in that example is the Non-LOSO-Unfiltered.

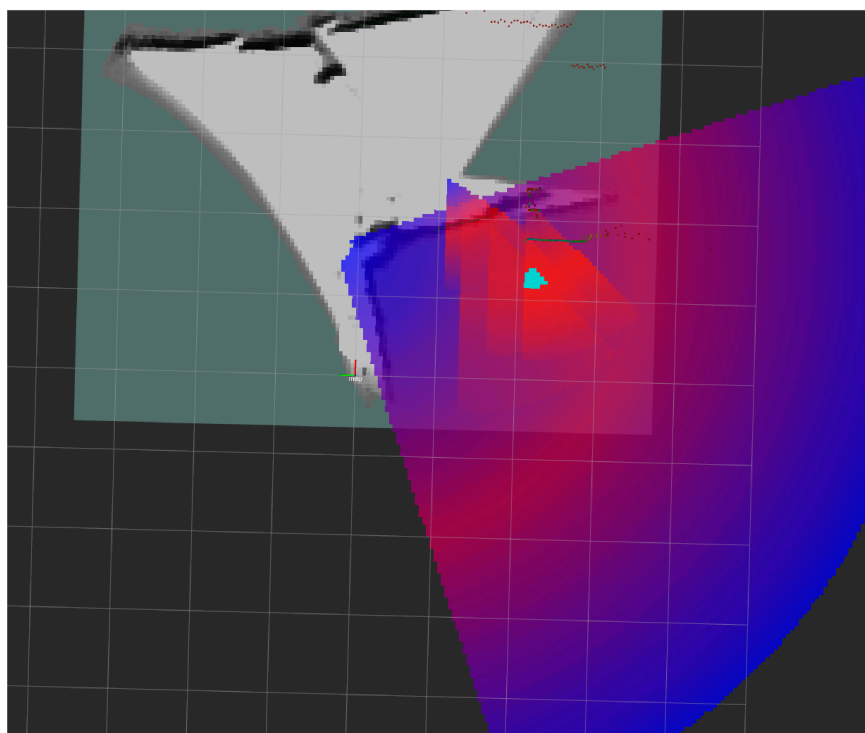


Figure 8. Results from the experiment. The model used in that example is the Non-LOSO-Filtered.

4. Discussion

This study set out to test whether a single hybrid CNN–ViT model can jointly perform through-wall victim detection and victim range estimation, and whether the resulting system generalizes to ranges and conditions beyond those seen during training. The results partially support this hypothesis. Under the LOSO protocol, the model reached an

F1-score of approximately 91–92% (0.92 ± 0.05 unfiltered, 0.91 ± 0.02 filtered) with only a few outliers, and on a held-out non-LOSO split it improved to 94% for the filtered data, indicating that presence detection shows partial robustness to inter-subject variability in body size, breathing pattern, and posture. Crucially, when the trained models were applied to data collected in a different environment and at longer ranges than the training set, classification performance was largely retained, based on only 33 external sessions (best external F1 $\approx 92\%$, for the Non-LOSO-Unfiltered), suggesting that the learned representation captures breathing-related radar signatures rather than environment-specific artifacts, but still the number of external sessions are not enough to conclude that the system would work on a range of distances and environments.

For range estimation, the proposed attention-rollout mechanism produced a mean absolute error of roughly 0.30–0.66 m under LOSO and 0.16–0.76 m on the externally collected data for all available models. These errors are encouraging given that the estimator is obtained directly from the transformer’s attention and requires no dedicated regression training. The Butterworth filtering stage had little effect on classification but consistently reduced the distance error, which is consistent with the interpretation that suppressing static clutter sharpens the temporal-range localization of the periodic breathing signature. Compared with our previous signal processing and machine-learning approaches [11,12], which estimate range from explicit features such as signal standard deviation or gradient-boosted regressors, the attention-based estimate trades a small amount of in-range precision for the ability to extrapolate to ranges not represented during training—an important property for search-and-rescue scenarios, where the victim’s distance is unknown a priori. We emphasize, however, that the attention overlay should be read as a localization cue rather than a faithful explanation of the classifier’s decision; the network integrates information beyond the attended region, and the attention map is therefore best interpreted qualitatively.

Several limitations temper these conclusions. The externally collected evaluation set is small (33 sessions), and the longer-range subset in particular contains few examples, so the generalization findings should be regarded as indicative rather than definitive. All measurements were obtained behind a single wall type and with a single subject per session; multi-victim scenarios, varied wall materials, and realistic debris were not examined.

Future work will address these points by expanding the dataset to more subjects, wall types, and cluttered environments by extending the framework to multi-target detection and separation; and directly comparing the attention-rollout estimator with trained regression baselines across the full range. We also intend to evaluate real-time, fully autonomous operation during continuous robot motion and to investigate sensor fusion with the platform’s existing LiDAR and camera streams to further improve localization confidence in operational search-and-rescue conditions.

Author Contributions: Conceptualization, A.-P.M. and G.N.; Methodology, A.-P.M., E.N.P. and G.N.; Software, A.-P.M. and E.N.P.; Validation, A.-P.M., E.N.P. and D.C.; Formal analysis, A.-P.M., E.N.P. and D.C.; Investigation, A.-P.M., E.N.P., G.N., S.A.M. and D.C.; Resources, A.-P.M. and S.A.M.; Data curation, A.-P.M. and E.N.P.; Writing—original draft, A.-P.M., E.N.P., G.N. and S.A.M.; Writing—review & editing, A.-P.M., G.N., S.A.M. and D.C.; Visualization, A.-P.M.; Supervision, G.N.; Project administration, S.A.M.; Funding acquisition, G.N., S.A.M. and D.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used in this research can be found at [12].

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Wang, D.; Yoo, S.; Cho, S.H. Experimental Comparison of IR-UWB Radar and FMCW Radar for Vital Signs. *Sensors* **2020**, *20*, 6695. [CrossRef] [PubMed]
- Chiou, M.; Epsimos, G.-T.; Nikolaou, G.; Pappas, P.; Petousakis, G.; Mühl, S.; Stolkin, R. Robot-Assisted Nuclear Disaster Response: Report and Insights from a Field Exercise. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022.
- Sieglwart, R.; Hutter, M.; Oettershagen, P.; Burri, M.; Gilitschenski, I.; Galceran, E.; Nieto, J. Legged and Flying Robots for Disaster Response. In Proceedings of the World Engineering Conference and Convention 2015 (WECC2015), Kyoto, Japan, 29 November–2 December 2015. [CrossRef]
- Liang, X.; Zhang, H.; Fang, G.; Ye, S.; Gulliver, T.A. An Improved Algorithm for Through-Wall Target Detection Using Ultra-Wideband Impulse Radar. *IEEE Access* **2017**, *5*, 22101–22118. [CrossRef]
- Liang, X.; Zhang, H.; Ye, S.; Fang, G.; Aaron Gulliver, T. Improved denoising method for through-wall vital sign detection using UWB impulse radar. *Digit. Signal Process.* **2018**, *74*, 72–93. [CrossRef]
- Liang, X.; Lv, T.; Zhang, H.; Gao, Y.; Fang, G. Through-wall human being detection using UWB impulse radar. *J. Wirel. Commun. Netw.* **2018**, *2018*, 46. [CrossRef]
- Uzunidis, D.; Kasnesis, P.; Patrikakis, C.Z.; Mitilineos, S.A. Machine Learning-Based Human Life Detection behind Walls Exploiting a UWB Radar Sensor. *Preprints* **2024**. [CrossRef]
- Paliodimos, E.; Papadopoulos, F.; Uzunidis, D.; Mitilineos, S.A. UWB radar dataset for victim detection through foliage in Search and Rescue operations. *Zenodo* **2024**. [CrossRef]
- Wang, W.; Du, N.; Guo, Y.; Sun, C.; Liu, J.; Song, R.; Ye, X. Human Detection in Realistic Through-the-Wall Environments using Raw Radar ADC Data and Parametric Neural Networks. *arXiv* **2025**, arXiv:2403.15468.
- Yang, S.; Kim, Y. Single 24-GHz FMCW Radar-Based Indoor Device-Free Human Localization and Posture Sensing With CNN. *IEEE Sens. J.* **2023**, *23*, 3059–3068. [CrossRef]
- Michalopoulos, A.-P.; Paliodimos, E.N.; Papadopoulos, F.; Nikolaou, G.; Patrikakis, C.; Mytilinaios, S.A. Victim Detection Using a Robot-Mounted UWB-Radar Platform. In Proceedings of the 2025 14th International Conference on Modern Circuits and Systems Technologies (MOCAS), Dresden, Germany, 11–13 June 2025; pp. 1–4. [CrossRef]
- Michalopoulos, A.-P.; Paliodimos, E.N.; Papadopoulos, F.; Nikolaou, G.; Patrikakis, C.Z.; Mytilinaios, S.A. A Dataset for Victim Detection in Search-and-Rescue Operations Using Robot-Mounted UWB-Radar Sensors. *Sci. Data* **2026**, *13*, 996. [CrossRef] [PubMed]
- Rohman, B.P.A.; Andra, M.; Nishimoto, M. Through-the-Wall Human Respiration Detection Using UWB Impulse Radar on Hovering Drone. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6572–6584. [CrossRef]
- Joseph, A.; Parmar, V.; Bagyaveereswaran, V. Design of Human Detection Robot for Natural calamity Rescue Operation. In Proceedings of the 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 22–23 March 2019; IEEE: New York, NY, USA, 2019; pp. 1–5. [CrossRef]
- Schroth, C.A.; Eckrich, C.; Kakouche, I.; Fabian, S.; von Stryk, O.; Zoubir, A.M.; Muma, M. Emergency Response Person Localization and Vital Sign Estimation Using a Semi-Autonomous Robot Mounted SFCW Radar. *IEEE Trans. Biomed. Eng.* **2024**, *71*, 1756–1769. [CrossRef] [PubMed]
- Zhao, Y.; Guendel, R.G.; Yarovoy, A.; Fioranelli, F. Distributed Radar-based Human Activity Recognition using Vision Transformer and CNNs. In Proceedings of the 2021 18th European Radar Conference (EuRAD), London, UK, 5–7 April 2022; pp. 301–304. [CrossRef]
- Wang, B.; Ding, C.; Chen, H.; Hong, H.; Zhu, X. Radar-based Human Activity Recognition with Range-distributed Time-Doppler Sparse Point Cloud and Multi-Channel PointNet. In Proceedings of the 2023 53rd European Microwave Conference (EuMC), Berlin, Germany, 19–21 September 2023; pp. 641–644. [CrossRef]
- Yataka, R.; Cardace, A.; Wang, P.P.; Boufounos, P.; Takahashi, R. RETR: Multi-View Radar Detection Transformer for Indoor Perception. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 19839–19869. [CrossRef]
- SensorLogicInc/modules. C. RM Young Company. 20 September 2025. Available online: <https://github.com/SensorLogicInc/modules> (accessed on 31 December 2025).
- ROBOTIS e-Manual LDS-01. ROBOTIS e-Manual. Available online: https://emanual.robotis.com/docs/en/platform/turtlebot3/appendix_lds_01/ (accessed on 26 July 2025).
- Takahashi, S.; Sakaguchi, Y.; Kouno, N.; Takasawa, K.; Ishizu, K.; Akagi, Y.; Aoyama, R.; Teraya, N.; Bolatkan, A.; Shinkai, N.; et al. Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. *J. Med. Syst.* **2024**, *48*, 84. [CrossRef] [PubMed]
- Suneetha, K.; KalaiPriya, O.; Sahoo, T. Optimizing Search and Rescue Missions Using Artificial Intelligence for Automated Aerial Image Processing and Analysis. In Proceedings of the 2025 15th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), Karachi, Pakistan, 20–21 December 2025; pp. 1–6. [CrossRef]

23. Czuba, A. Vision Transformer for Classification of UAV and Helicopters Using Micro-Doppler Spectrograms in Surveillance Radar. *arXiv* **2025**, arXiv:2512.00374.
24. Nguyen, N.T.; Do, T.N. Temporal Micro-Doppler Spectrogram-based ViT Multiclass Target Classification. *arXiv* **2025**, arXiv:2511.11951.
25. Zhang, T.; Fan, Y. A 3D U-Net Based on a Vision Transformer for Radar Semantic Segmentation. *Sensors* **2023**, *23*, 9630. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.