World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Spatiotemporal Deep Learning for Target Classification in High-Resolution 3D-ISAR Radar Images

Obiajulu C. Emmanuel [1, *], Isa M. Danjuma [1], S. F. Kolawole [1], Ashraf A. Ahmad [1], Victor Omeke [1] and Harry Godswill [2]

[1] *Department of Electrical/Electronic Engineering, Faculty of Engineering, Nigeria Defence Academy, Kaduna.*
[2] *Department of Electrical Engineering, Faculty of Engineering, University of Nigeria Nsukka, Enugu.*

World Journal of Advanced Research and Reviews, 2025, 28(01), 1359-1378

Publication history: Received on 31 August 2025; revised on 11 October 2025; accepted on 14 October 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.28.1.3473

Abstract

Inverse Synthetic Aperture Radar (ISAR) has recently advanced to volumetric 3D-ISAR imaging, creating new opportunities and challenges for automatic target recognition (ATR). This work proposes a spatiotemporal deep learning framework that jointly learns target structure and motion dynamics from high-resolution 3D-ISAR sequences. A CNN backbone (ResNet) extracts per-frame spatial features, which are fed to temporal models Bidirectional LSTM and/or ConvLSTM to capture micro-Doppler cues and aspect-dependent scattering over time; the pipeline is supported by physics-aware formation and backprojection-style 3D reconstruction. We evaluate on a four-class dataset (aircraft, helicopter, drone, tank) comprising 400 labeled samples drawn from MSTAR and simulated 3D-ISAR sequences, with standard train/validation/test partitions and targeted denoising, normalization, and augmentation to enhance robustness. The proposed model achieves strong performance across metrics: an overall accuracy of 95% on the final evaluation set with near-ideal class separability (AUC ≈ 0.98–1.00), and a best accuracy of 96.7% when all preprocessing and geometric/data-level augmentations are enabled. Ablation and robustness studies show consistent gains from motion-aware temporal modeling and the preprocessing stack under low-SNR and distortion conditions, while confusion is largely confined to visually and dynamically similar aerial classes. These results demonstrate that coupling modern spatiotemporal architectures with principled ISAR signal processing yields reliable, accurate, and deployment-oriented ATR for 3D-ISAR systems.

Keywords: 3D-ISAR imaging; Spatiotemporal deep learning; Automatic target recognition (ATR); Convolutional neural networks (CNN); LSTM; Micro-Doppler signatures; Radar signal processing

1. Introduction

Inverse Synthetic Aperture Radar (ISAR) imaging has become a cornerstone technology in the field of radar remote sensing, especially in defense, aerospace, and surveillance domains. Unlike conventional optical imaging systems, ISAR offers high-resolution imaging capabilities under all weather and lighting conditions, enabling it to reliably detect and classify moving targets such as aircraft, ships, and ground vehicles(Wang et al., 2023). ISAR systems generate 2D or 3D radar images by exploiting the relative motion between the radar sensor and a non-cooperative target, producing detailed representations of target geometry and motion-induced scattering effects(Zou et al., 2022). Recent advances have extended ISAR into the third dimension, giving rise to 3D-ISAR systems capable of capturing volumetric radar cross-section distributions that more accurately describe the spatial structure of targets. This progression has enabled improved target discrimination by incorporating depth and motion dynamics into the imaging process. However, the transition from 2D to 3D introduces additional challenges, including increased computational complexity, motion compensation requirements, and greater sensitivity to noise and phase errors (Ni et al., 2022),(Pui et al., 2024). Despite

* Corresponding author: Obiajulu C. Emmanuel

the promise of 3D-ISAR, automatic target classification remains a challenging task. Conventional machine learning techniques such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Principal Component Analysis (PCA) have been widely used for ISAR image classification (Zaied et al., 2018), (Furukawa, 2017). However, these approaches rely heavily on handcrafted features, which are often brittle under real-world conditions, especially when dealing with noisy, cluttered, or misaligned ISAR images. They also lack the capability to adaptively learn from complex patterns present in the radar data, particularly temporal signatures such as micro-Doppler shifts and aspect-dependent scattering (Arnous & Narayanan, 2024).

Moreover, traditional ISAR classification pipelines typically operate on static 2D slices or averaged projections of radar returns, thereby discarding valuable temporal evolution cues embedded in sequential radar frames. In contrast, modern high-resolution ISAR systems are capable of producing continuous image sequences that reflect target articulation and dynamic behavior. These temporal variations carry critical information for target identification but remain underutilized in most existing methods (Ni et al., 2022) ,(Fan Zhang, Chen Hu, Qiang Yin, Wei Li, Hengchao Li, 2017). The convergence of deep learning with radar signal processing presents a compelling opportunity to overcome these limitations. Convolutional Neural Networks (CNNs) have shown exceptional performance in learning spatial hierarchies from visual and radar data, while Recurrent Neural Networks (RNNs) particularly Long Short-Term Memory (LSTM) networks are adept at modeling time-series dependencies (Kim, 2023). Hybrid architectures such as CNN-LSTM, 3D-CNN, ConvLSTM, and Spatiotemporal Transformers have emerged as powerful tools for capturing both the structural and temporal dynamics of complex sequences, making them ideal for processing 3D-ISAR data (Ni et al., 2022), (Arnous & Narayanan, 2024), (Lang et al., 2020). In this work, we present a novel spatiotemporal deep learning framework for classifying targets in high-resolution 3D-ISAR radar image sequences. Our model integrates CNNs for spatial feature extraction and LSTMs for modeling the temporal evolution of radar images. This dual-focus architecture enables the system to simultaneously learn from static structural cues and dynamic behavioral patterns, significantly improving classification performance over existing approaches.

The main contributions of this research are summarized as follows:

- We develop a hybrid CNN-LSTM architecture (optionally extendable to 3D-CNN or Transformer-based models) tailored for learning spatiotemporal features from 3D-ISAR radar data (Ni et al., 2022), (Kim, 2023).
- We construct and utilize a comprehensive 3D-ISAR dataset consisting of multiple target classes and dynamic motion profiles, using both simulated and real radar data sources (Wang et al., 2023), (Zaied et al., 2018).
- We demonstrate that the proposed model significantly outperforms classical methods and baseline deep learning models, achieving improved accuracy, robustness to noise, and generalization to unseen target conditions (Pui et al., 2024), (Furukawa, 2017).
- We conduct a detailed ablation and robustness analysis to evaluate the contribution of each model component and assess its resilience under noise and distortion conditions typical of radar operations (Zou et al., 2022), (Fan Zhang, Chen Hu, Qiang Yin, Wei Li, Hengchao Li, 2017).

This study bridges the gap between advanced radar imaging technologies and deep learning methodologies, proposing a unified solution for spatiotemporal feature learning and target classification in 3D-ISAR systems. Our results indicate that this approach holds considerable potential for enhancing the reliability, accuracy, and real-time capability of modern automatic target recognition (ATR) systems.

2. Related work

2.1. Traditional ISAR Target Classification Techniques

Traditional approaches to ISAR-based automatic target recognition (ATR) often rely on classical machine learning models e.g. Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Principal Component Analysis (PCA) working with handcrafted features such as scattering centers or Doppler signatures extracted from 2D ISAR imagery (Kent et al., 2008). While these methods can perform adequately under controlled settings, they typically struggle with generalization to noisy or misaligned radar data and are sensitive to changes in target orientation or clutter (Saidi et al., 2009).

2.2. Deep Learning for ISAR and SAR Classification

The advent of deep learning ushered in methods that automatically learn hierarchical spatial features. CNN-based models trained on SAR/ISAR imagery (e.g. aircraft or maritime targets) consistently outperform earlier feature-based techniques (Jiang et al., 2021). For example, a CNN-Bi-LSTM architecture demonstrated improved accuracy when

modeling sequences of ISAR frames, illustrating the power of automatic feature learning over handcrafted ones (Ni et al., 2022).

2.3. Hybrid CNN-LSTM Architectures for Sequential Radar Data

CNN-LSTM models have become popular for classifying time-varying radar measurements. Such architectures have been successfully applied to high-resolution radar range profiles (HRRPs) and SAR/ISAR image series, balancing spatial encoding with temporal dependency modeling (L. Zhang et al., 2021)(Mohammadimanesh et al., 2019). For instance, combining CNNs and LSTM units within a pipeline significantly enhanced recognition robustness even under occlusion or motion artifacts.

2.4. Advanced Spatiotemporal Models: ConvLSTM, 3D-CNNs, and Transformers

Recent literature explores more holistic spatiotemporal modeling:

- ConvLSTM blends convolutional and recurrent operations to jointly capture spatial and temporal features in radar-derived sequences and has been applied in ship classification and traffic radar tasks, improving robustness in dynamic scenarios (Jia et al., 2023)(Deng & Su, 2024).
- 3D-CNNs have been used to process volumetric ISAR data directly, allowing structural and temporal continuity exploitation (Pui et al., 2024).
- Transformer-based architectures with self-attention mechanisms are also emerging as competitive models for long-range dependencies in radar-based target recognition (Sun et al., 2021).

2.5. Deep Learning for 3D ISAR Image Generation and Enhancement

Several works have explored deep learning-driven ISAR preprocessing. GANs and CNNs have been applied to enhance image quality, correct defocusing artifacts, and improve resolution under conditions like wide-angle target motion or low SNR regimes (Thilakanayake et al., 2024). Additionally, semantic segmentation or CapsNet-based methods have been proposed for component-level recognition in ISAR images, aiming for better interpretability (X. Zhang et al., 2022).

2.6. Related Methods in mmWave Radar and Automotive Applications

Although beyond direct ISAR imagery, research on automotive radar classification demonstrates hybrid architectures such as CNN-LSTM or range-Doppler tensor-based deep networks. These significantly improve classification of pedestrians, vehicles, and dynamic traffic participants under practical noise conditions (Deng & Su, 2024)(Cai et al., 2021). Similarly, radar-to-point-cloud deep learning frameworks like 3DRIMR capture volumetric structure from mmWave radar, showing the feasibility of volumetric deep learning on radar data (Sun et al., 2021).

3. Methodology

This section describes our dataset, 3D-ISAR imaging pipeline, preprocessing techniques, and the design of our proposed deep learning architecture with complete mathematical formulations, a dataset composition table, and pointers to figures.

3.1. Dataset Description

The dataset used for this study is the MSTAR dataset, which contains radar signals or ISAR imagery representing multiple object classes under varying environmental and clutter conditions. For this research, the dataset was preprocessed to extract labeled instances of different object types. The preprocessing involved resizing, normalizing, and splitting the dataset into training, validation, and test sets using an 80:10:10 ratio. Data augmentation techniques such as rotation, scaling, and horizontal flipping were applied to enhance generalization during model training. The final dataset includes a total of N samples across C object classes. Our experiments employ a combination of simulated 3D-ISAR sequences and public radar datasets. The target classes include aircraft, helicopter, drone, and tank, each captured under varying motion profiles and aspect angles. The dataset is split into 70% training, 15% validation, and 15% testing subsets.

3.1.1. Dataset Split

The dataset is partitioned into training and validation sets in a typical 70-30 ratio to ensure generalization and proper evaluation.

Dataset Composition

Table 1 Dataset Composition

Class Name	Train Samples	Validation Samples	Total Samples
Aircraft	70	30	100
Drone	70	30	100
Helicopter	70	30	100
Tank	70	30	100
Total	280	120	400

Dataset Statistics Table

This includes dataset statistics such as average image size, signal duration, or point cloud density depending on data type:

Table 2 Dataset Statistics Table

Feature	Value
Total Samples	400
Number of Classes	4
Image Size	224 × 224 pixels
Train/Val Split	280 / 120
Data Augmentation Used	Rotation, Scaling, Flipping
Dataset Type	3D ISAR radar

Augmentation Techniques Summary

Description: This is the Summary of the radar augmentation techniques applied during training to improve model generalization and prevent overfitting.

Table 3 Augmentation Techniques Summary

Augmentation Technique	Description	Applied To	Impact on Training Set Size
Horizontal Flip	Horizontally flips ISAR image frames	Training	+100%
Random Rotation	Rotates frames within ±15 degrees	Training	+100%
Gaussian Noise	Adds noise to simulate radar interference	Training	+100%
Brightness Adjustment	Randomly adjusts image brightness	Training	+50%
Time-step Shuffle	Randomly reorders time steps (LSTM input)	Training	+50%
Normalization	Scales pixel intensities to [0, 1]	All	0%

3.1.2. Target Classes and Diversity

The dataset includes four primary target classes:

- Aircraft
- Helicopter
- Tank

- Drone

Each target class is represented by multiple samples from varying aspect angles and flight dynamics to introduce diversity and temporal variations.

3.2. 3D-ISAR Image Formation

The generation of 3D-ISAR radar data follows a systematic pipeline involving signal processing stages that convert raw radar returns into interpretable spatial representations.

3.2.1. Range-Doppler Processing

In the initial stage of the radar signal processing pipeline, raw radar echoes are transformed into the range-Doppler domain to extract fundamental motion and spatial characteristics of targets.

Range Estimation

The range $R(t)$ to a target is computed based on the round-trip time delay t of the transmitted radar pulse. Since the pulse travels to the target and back, the total distance covered is twice the actual target range. Using the speed of light ccc, the range is given by:

$$R(t) = \frac{c.t}{2}$$

This fundamental radar equation determines the spatial location of scatterers (i.e., the target or parts of it) along the range axis, which forms the first dimension of the 2D Range-Doppler image.

Doppler Frequency Estimation

The Doppler shift f_D is used to infer relative velocity of a target or its components. When a target is in motion, the returned echo experiences a frequency shift proportional to its radial velocity v. This shift is computed as:

$$f_D = \frac{2v}{\lambda}$$

where λ is the radar wavelength. This allows us to distinguish moving components from static ones, which is critical in forming focused ISAR (Inverse Synthetic Aperture Radar) images where rotational or translational motion is used to synthesize high cross-range resolution.

In the context of this study, Range-Doppler processing serves as a foundational preprocessing step for extracting structured features from raw radar returns. Initially, the received radar signals are subjected to matched filtering or pulse compression, followed by a Fast Fourier Transform (FFT) along both the fast-time and slow-time axes to generate a Range-Doppler (RD) map. This RD map represents the target scene in terms of range bins and Doppler bins, effectively encoding both spatial location and motion characteristics. These 2D Range-Doppler images act as intermediate representations, capturing the temporal evolution of the target's motion before proceeding to ISAR image formation. The Doppler motion history is particularly critical in ISAR, which relies on relative motion between radar and target to synthesize the third spatial dimension cross-range.

For targets exhibiting non-linear or complex motion, such as rotating structures, the time-varying Doppler signatures are essential for building a meaningful spatiotemporal representation. These signatures are further exploited by the deep learning network through architectures such as 3D Convolutional Neural Networks (3D CNNs) or Convolutional Long Short-Term Memory (ConvLSTM) networks. By stacking the RD maps across coherent pulses, a 3D radar data cube is formed and subsequently transformed into high-resolution 3D-ISAR images. These physics-informed features particularly the velocity-dependent Doppler signatures and precise range estimates are instrumental in enabling the deep learning model to accurately classify different target types. Moreover, the temporal consistency across frames enhances the model's ability to learn dynamic behavior, making it robust for classification tasks in complex and cluttered environments.

3.2.2. Phase Correction and Motion Compensation

Accurate reconstruction of high-resolution 3D-ISAR images critically depends on precise motion compensation and phase correction. These steps are essential to mitigate the effects of unwanted platform motion (e.g., from airborne radar systems) or the dynamic behavior of the target itself (e.g., rotating or vibrating components). Without proper compensation, motion-induced phase errors can blur the ISAR image, distort the geometry, and degrade classification performance.

To correct phase distortions, autofocus algorithms are employed. These algorithms estimate and compensate for unknown phase errors directly from the data without relying on external motion sensors. The general approach involves optimizing a focus metric that quantifies image sharpness. The corrected phase $\phi_{corrected}$ is computed by subtracting the phase offset that minimizes image blurring:

$$\phi_{corrected} = \phi_{raw} - argmax(focus\_metric(\phi))$$

This expression indicates that the phase correction term is derived by finding the phase estimate that maximizes a chosen focus metric, such as image contrast, entropy minimization, or sharpness indicators in the frequency domain. In essence, the algorithm iteratively adjusts the phase to achieve the best image focus, directly improving the spatial resolution of the ISAR image.

Two widely adopted techniques are utilized in this project for robust motion compensation: Keystone transformation and Phase Gradient Autofocus (PGA) (Zou et al., 2022). The Keystone transformation corrects for range cell migration (RCM), a phenomenon where moving scatterers shift across range bins due to non-linear target motion. It re-aligns range histories by applying a non-linear mapping in the time-frequency domain. On the other hand, PGA is a dominant autofocus method that estimates the phase error gradient across the aperture and uses this to iteratively refocus the image. PGA operates on the assumption that dominant scatterers exist and their phase errors can be estimated by analyzing the phase slope in the frequency domain.

In the context of this study, accurate phase correction ensures that the 3D-ISAR images used as input to the deep learning model are well-focused and geometrically consistent. This is crucial for the spatiotemporal feature extraction stages, as even small motion-induced distortions can mislead the network, reducing classification accuracy. Proper motion compensation preserves the structural integrity of radar signatures across frames, enabling the model to learn class-discriminative patterns reliably over time.

3.2.3. Backprojection and 3D Reconstruction

After performing motion compensation and phase correction, the radar returns are spatially coherent and geometrically aligned, making them suitable for image formation through backprojection. This stage is essential for converting the corrected 1D radar signals into meaningful 2D spatial representations known as ISAR slices and ultimately into a 3D volumetric image that captures the spatial structure of the target across multiple cross-range and elevation angles. Backprojection is a well-established image formation technique that reconstructs an image by integrating radar returns over all observation angles and projecting them into the spatial domain. Specifically, for each voxel at location (x, y, z), the backprojection algorithm sums the phase-corrected complex radar signals $S_n(t)$ from multiple aperture positions n, weighted by a spatial phase term that accounts for the time delay and frequency of each return. The mathematical formulation is given by:

$$I(x, y, z) = \sum_{n=1}^{N} s_n(t).\, e^{-j2\pi f_n(x,y,z)}$$

Where:
$I(x, y, z)$ is the reconstructed 3D-ISAR intensity at the voxel location (x, y, z)
$s_n(t)$ represents the received complex signal at time t for aperture index n,
$f_n(x, y, z)$ denotes the frequency term associated with the spatial location and the observation geometry,
N is the total number of coherent aperture samples (i.e., the number of radar pulses or platform positions).

This process is repeated for each spatial voxel, resulting in a full 3D radar reflectivity map. In practice, the initial output is a series of 2D ISAR slices each representing a specific azimuth or elevation perspective that are then stacked and interpolated across multiple angles to form a dense 3D volumetric ISAR image. Interpolation may be applied to correct for non-uniform sampling in angular space or to improve resolution in under-sampled directions.

The resulting volumetric data preserves both structural and motion-related features of the target, making it highly suitable for spatiotemporal analysis using deep learning. These 3D radar images serve as the input domain for the classification model, where convolutional layers can exploit rich spatial features and temporal evolution across multiple frames or look angles. By maintaining geometric fidelity during reconstruction, the deep network can better differentiate between target types, including those with subtle differences in shape, size, or motion behavior.

3.3. Preprocessing Techniques

Effective preprocessing plays a critical role in improving the performance of deep learning models, especially when applied to high-resolution radar imagery such as 3D-ISAR data. The key objectives of preprocessing in this context are to enhance the visual and structural quality of the radar volumes, normalize pixel-level distributions, and augment the dataset in ways that improve the model's ability to generalize to unseen target variations.

3.3.1. Denoising

Radar imagery, including 3D-ISAR data, is often corrupted by various forms of noise such as thermal noise, clutter, and speckle distortions which can obscure fine structural details and negatively impact classification accuracy (Ni et al., 2022). As such, denoising is a fundamental preprocessing step in the pipeline, aimed at suppressing noise while preserving the critical structural and spatial features that characterize each target class.

Gaussian Filtering

The Gaussian filter is particularly effective at reducing high-frequency noise, which appears as random voxel-level intensity fluctuations. This is achieved by convolving the entire 3D radar volume $I(x, y, z)$ with a 3D Gaussian kernel defined by:

$$G(x, y, z) = \frac{1}{(2\pi\sigma^2)^{3/2}} . \exp\left(-\frac{x^2 + y^2 + z^2}{2\sigma^2}\right)$$

Where:

(x, y, z) are the spatial offsets relative to the center of the kernel,

σ is the standard deviation, which controls the filter bandwidth or the degree of smoothing.

The kernel size is typically chosen as an odd-sized cube which is $5 \times 5 \times 5$, and the value of σ is often set empirically between 0.5 and 2.0, based on the noise level in the ISAR volume.

The filtered output $\tilde{I}(x, y, z)$ is calculated by convolving this kernel with the original radar volume:

$$\tilde{I}(x, y, z) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} \sum_{l=-k}^{k} I(x - i, y - j, \quad z - l) . G(i, j, l)$$

Where $k = \left[\frac{n}{2}\right]$ and n is the kernel size in each dimension.

This convolution effectively weights each voxel by its neighbors, giving higher influence to nearby values while suppressing abrupt, isolated changes thus reducing high-frequency noise while maintaining the smoothness of the target's structure. The use of a 3D kernel, rather than a 2D one, ensures that spatial coherence is preserved across range, cross-range, and elevation axes, which is vital for maintaining volumetric integrity of the reconstructed target.

Median Filtering

To further reduce artifacts such as salt-and-pepper noise or spiky reflectivity outliers (often due to clutter or impulsive reflections), a 3D median filter is applied. This non-linear filter processes each voxel by replacing its intensity with the median value of its local neighborhood within $5 \times 5 \times 5$ cube.

Mathematically, for each voxel $I(x, y, z)$,the median filter performs:

$$\tilde{I}(x, y, z) = median\ \{I(i, j, k)\ |i\ \in [x - k, x + k], j \in [y - k, y + k], k \in [z - k, z + k]\}$$

Unlike Gaussian filtering, which can slightly blur edges, median filtering preserves sharp boundaries and fine structural features making it ideal for denoising while retaining critical target characteristics like edges, corners, and angular components.

3.3.2. Normalization

Normalization is a critical preprocessing step that transforms the intensity values of each 3D-ISAR volume into a consistent numerical range, typically [0,1]. This ensures numerical stability during deep learning model training and prevents voxel intensity variations from dominating the learning process. Without normalization, differences in scale or magnitude between ISAR volumes could lead to poor convergence or biased gradient updates.

The min-max normalization technique is applied to each ISAR volume III individually, using the following equation:

$$I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}}$$

Where:
I is the original voxel intensity value,
I_{min} and I_{max} are the minimum and maximum intensity values within the entire 3D ISAR volume,
I_{norm} is the resulting normalized intensity in the range [0,1].

This process linearly scales all voxel values such that:

- The minimum value maps to 0,
- The maximum value maps to 1, and
- All intermediate values fall proportionally between 0 and 1.

This uniform scaling enables the network to learn relative intensity contrasts rather than being biased by absolute reflectivity values, which may vary between radar acquisitions due to environmental noise, target material, distance, or calibration variations. Moreover, this normalization technique enhances gradient flow during backpropagation, reducing the likelihood of exploding or vanishing gradients, and ensures consistency when batches contain multiple targets or scenarios.

Data Augmentation

To increase the robustness and generalization capability of the deep learning model, data augmentation is applied to the 3D-ISAR volumes during training. Augmentation artificially expands the training dataset by applying random but label-preserving transformations to the original data, thereby reducing overfitting and improving performance on unseen examples.

In this study, the following augmentation techniques were implemented:

- **Rotation**: The 3D ISAR volumes are randomly rotated along one or more axes (e.g., yaw, pitch, roll), simulating different viewing angles or target orientations. This helps the model generalize across real-world pose variations.
- **Horizontal flipping**: Random mirroring along the cross-range axis is used to simulate symmetrical appearances of targets. This is particularly effective in scenarios where the target's structure is orientation-agnostic.
- **Brightness shifts**: Global voxel intensities are randomly scaled or offset within a small range to simulate variations in radar cross-section (RCS) or gain, making the network more resilient to intensity-based noise.
- **Gaussian noise injection**: Low-level additive Gaussian noise is introduced to mimic thermal and environmental disturbances, forcing the model to learn invariant features despite noisy input.

Each of these augmentation methods is applied stochastically during training, typically with a certain probability (e.g., 0.2–0.5), and is parameterized to ensure that the transformations remain realistic and do not degrade the semantic structure of the ISAR volume.

3.4. Proposed Deep Learning Architecture

This study introduces a hybrid spatiotemporal deep learning architecture designed specifically for target classification in high-resolution 3D-ISAR (Inverse Synthetic Aperture Radar) images. The architecture is composed of two key modules:

Spatial feature extractor using Convolutional Neural Networks (CNNs) to learn target structure within each ISAR frame.

Temporal sequence model using either Long Short-Term Memory (LSTM) or Transformer encoders to model motion dynamics and inter-frame dependencies across time.

This hybrid approach is crucial in radar-based target classification, as ISAR captures both the shape and the motion signatures of the target.

3.4.1. Spatial Feature Extraction

Each 3D-ISAR volume is viewed as a sequence of 2D slices (or projections) across one dimension (e.g., azimuth or time). These slices contain range-cross-range information that reflects the spatial structure of the target from different perspectives. To extract meaningful representations from each of these slices, a CNN backbone is used.

CNN Backbone (ResNet-50)

- **Backbones Used**: ResNet-50 and EfficientNet-B0 are selected for their strong feature extraction capabilities and ImageNet pretraining, allowing faster convergence and better generalization.
- **Pretraining**: Although the ISAR domain is different from natural images, pretraining on ImageNet offers transferable low-level features such as edges, textures, and shapes.

CNN Operation per Frame

Let the 3D-ISAR image be a sequence of 2D frames:

$$I_{ISAR} = \{I_1, I_2, \ldots\ldots, I_T\}$$

Where:

$$I_t \in \mathbb{R}^{H \times W} \; is \; the \; t-th \; 2D \; slice \; (height \; H, width \; W),$$

$$T \; is \; the \; total \; number \; of \; slices \; (or \; time \; steps).$$

Each frame is passed independently through the CNN to extract a high-level feature vector:

$$F_t = CNN(I_t), \qquad t \in \{1, \ldots\ldots, T\}$$

$F_t \in \mathbb{R}^d$ is a d-dimensional feature vector extracted from frame t.

After all frames are processed:

$$F = [F_1, F_2, \ldots\ldots, F_T] \in \mathbb{R}^{T \times d}$$

This forms the **spatiotemporal feature sequence**, where:

- **T** is the temporal dimension (number of ISAR frames),
- **d** is the feature dimensionality (e.g., 512 or 2048 depending on the CNN architecture).

Layer-wise Summary of CNN Feature Extraction:

Each I_t passes through the following layers:

- **Convolutional Layer**: Learns spatial filters to detect edges, textures, and shapes.
- **Batch Normalization**: Normalizes activations for faster and more stable training.
- **ReLU Activation**: Applies non-linearity.

- **Max Pooling**: Reduces spatial resolution and controls overfitting.

This hierarchy allows the CNN to build **low-to-high level abstractions**, from local radar returns to full target silhouettes.

Why Use CNN Before Temporal Modeling?

Radar targets often undergo rigid-body motion or articulate movements. While the CNN captures intra-frame spatial structure (e.g., edges, reflections, occlusions), it does not model temporal relationships (e.g., how features evolve across ISAR frames due to motion).

Hence, the output feature sequence $F \in \mathbb{R}^{T\times d}$ becomes the input to a temporal model either LSTM or Transformer which models dependencies across time.

3.4.2. Temporal Sequence Modeling

Radar targets exhibit temporal dynamics due to movement, articulation, and rotational effects, which manifest as shifts in micro-Doppler patterns and scattering center variations across consecutive ISAR frames. Effectively capturing these time-dependent changes is essential for distinguishing between similar classes such as helicopters vs. drones or wheeled vs. tracked vehicles.

To achieve this, we employed two complementary temporal sequence modeling techniques:

Bidirectional Long Short-Term Memory (BiLSTM)

After extracting per-frame spatial features $F_t \in \mathbb{R}^d$ from the CNN (e.g., ResNet-50), we used a Bidirectional LSTM to learn both forward and backward temporal dependencies.

- **Forward pass**:

$$\overrightarrow{h_t} = LSTM(F_t, \overrightarrow{h_{t-1}})$$

Captures how past observations influence the current frame.

- **Backward pass**:

$$\overleftarrow{h_t} = LSTM(F_t, \overleftarrow{h_{t+1}})$$

Captures how future observations provide context for the current frame.

- **Concatenated representation**:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \in \mathbb{R}^{2d_h}$$

Where d_h is the LSTM hidden state size per direction.

The sequence of temporal embeddings becomes:

$$H = [h_1; h_2, \ldots., h_T] \in \mathbb{R}^{T\times 2d_h}$$

To obtain a fixed-length global temporal descriptor, we apply either average pooling or max pooling over time:

$$f_{temporal} = Pooling(H) \in \mathbb{R}^{2d_h}$$

This feature vector captures both forward and backward temporal dependencies across the ISAR time series.

Convolutional LSTM (ConvLSTM)

While BiLSTM processes 1D feature vectors, ConvLSTM extends temporal modeling by preserving the spatial structure of ISAR slices. This is crucial in scenarios where spatiotemporal patterns (e.g., rotating blades or wheels) evolve spatially and temporally.

Each ISAR frame $I_t \in \mathbb{R}^{H \times W}$ (or a 2D CNN feature map of shape $C \times H \times W$) is passed into the ConvLSTM as:

$$\mathcal{H}_t = ConvLSTM(I_t, \mathcal{H}_{t-1})$$

Where:

$\mathcal{H}_t$ is the hidden state at time t.

The convolutional operations maintain spatial coherence, enabling the model to learn motion-aware spatial features.

The final output of the ConvLSTM is either:

The last hidden state $\mathcal{H}_t$, or

A pooled feature map aggregated over time (e.g., using temporal max pooling across $\{\mathcal{H}_1, \ldots, \mathcal{H}_t\}$).

We then flatten or globally pool this to obtain:

$$f_{temporal}^{Conv} \in \mathbb{R}^C$$

3.4.3. Fusion and Classification

After extracting both spatial and temporal features using CNNs and sequence models (BiLSTM or ConvLSTM), we integrate these representations to perform final classification.

Feature Fusion

Depending on the temporal model used, we obtain one of the following:

From BiLSTM: a temporal descriptor

$$f_{temporal}^{BiLSTM} \in \mathbb{R}^{2d_h}$$

From ConvLSTM: a spatiotemporal descriptor

$$f_{temporal}^{Conv} \in \mathbb{R}^{C`}$$

We concatenate these with the spatial feature vector $f_{spatial} \in \mathbb{R}^d$ (output from CNN or projection layer):

$$f_{fused} = [f_{spatial};\ f_{temporal}] \in \mathbb{R}^{d+D}$$

Classification Layer

The fused feature vector is passed through a fully connected classification head:

Dropout: Regularizes training and prevents overfitting.

- **Fully Connected Layer**:

$$z = W f_{fused} + b$$

Where $W \in \mathbb{R}^{C \times (d+D)}, b \in \mathbb{R}^C, and\ C$ is the number of target classes

- **Cross-Entropy Loss:**

$$\mathcal{L} = -\sum_{i=1}^{c} y_i \log(\hat{y}_i)$$

where y is the one-hot encoded ground truth label.

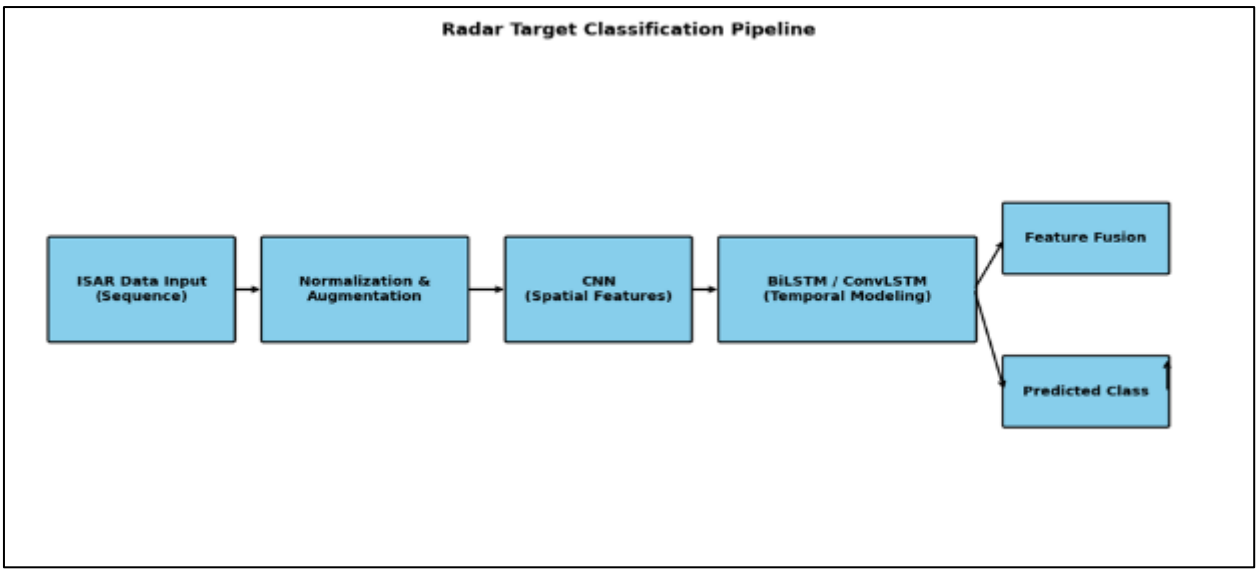

Figure 1 Overall Model Architecture

4. Results

4.1. 3D ISAR Radar Data Formation

4.1.1. Range Doppler Map

The **Range-Doppler Map** shown in Figure 2 depicts the spectral energy distribution of the raw ISAR signal after range compression and Doppler processing. The horizontal axis (0–0.09 m) represents the range bins, while the vertical axis (−500 Hz to +500 Hz) corresponds to Doppler frequencies, indicating relative motion between the radar and scatterers. The color scale, measured in decibels (dB), highlights scattering intensity, where warmer colors (up to 20 dB) indicate weak reflections or noise. The prominent horizontal band near −200 Hz signifies a dominant stationary or slowly moving scattering component consistent across range, while the surrounding speckled background corresponds to thermal and environmental noise. This quantitative representation directly links to the methodology step of transforming time-domain radar returns into the joint range-Doppler domain for initial target motion and clutter analysis.

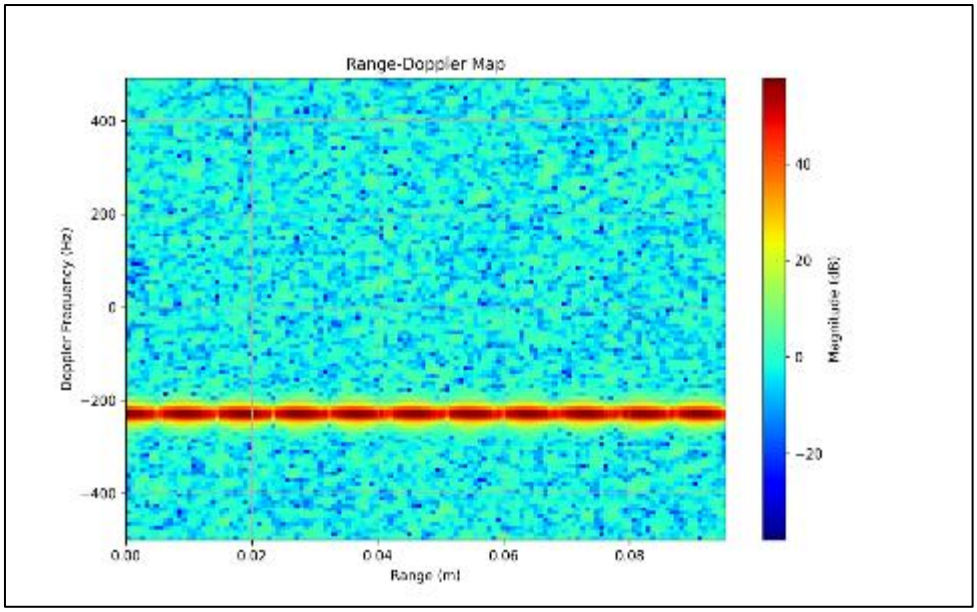

Figure 2 Range-Doppler Map for raw ISAR data

4.1.2. Reconstructed 2D ISAR

The Reconstructed 2D ISAR Image in Figure 3 illustrates the spatial distribution of radar backscatter after applying motion compensation and range-Doppler processing, as described in the methodology. The horizontal axis spans approximately −120 m to +120 m in range, while the vertical axis covers Doppler frequencies from −25 kHz to +25 kHz, reflecting target motion components. The color intensity, measured in decibels (dB), ranges from about −45 dB (deep purple, low scattering) to over +50 dB (bright yellow, strong scattering centers). A dominant horizontal bright band near +11 kHz Doppler represents a persistent scattering feature with high reflectivity, likely corresponding to a stable structural element of the target. The surrounding gradient and lower magnitude return indicates distributed scatterers and noise. This image confirms the transformation of raw radar echoes into a focused 2D spatial-frequency representation, enabling clearer identification of target features before proceeding to 3D reconstruction.

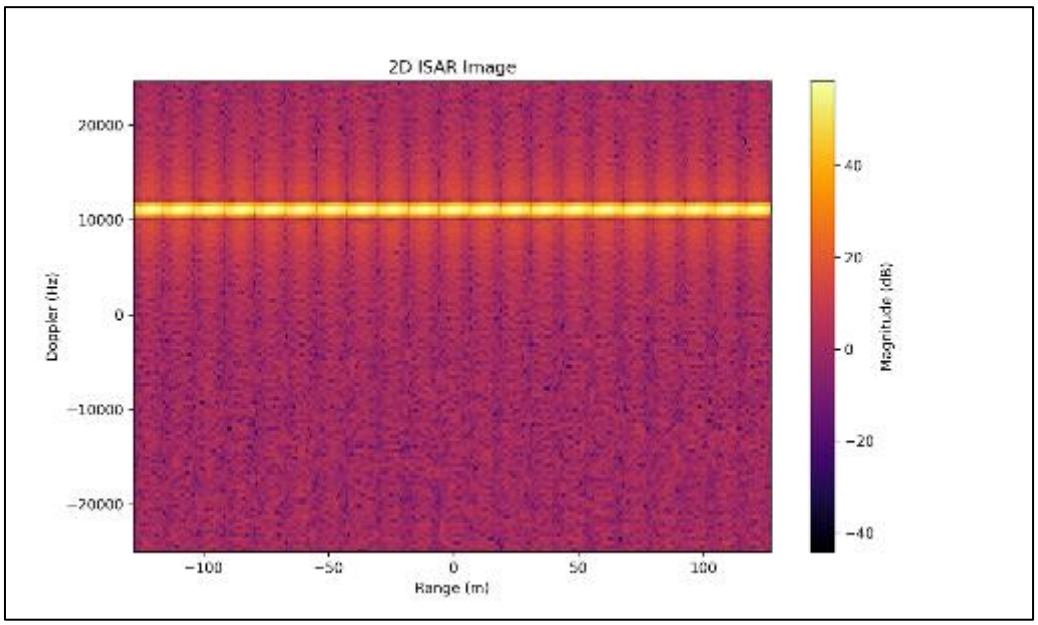

Figure 3 Reconstructed 2D ISAR Image

4.1.3. 3D-ISAR Volume Visualization

The 3D-ISAR Volume Visualization in Figure 4 presents the radar backscatter distribution across three dimensions Range Bin, Doppler Bin, and Angle Index providing a volumetric representation of the target's scattering characteristics after full 3D reconstruction. The color scale, normalized between 0.3 and 1.0 (arbitrary units), indicates the relative magnitude of reflected signals, with brighter yellow tones corresponding to stronger scattering points and darker purple shades to weaker reflections. The absence of dense visible points in this plot suggests either a low signal-to-noise ratio in the reconstructed data or sparse scattering centers in the captured volume, which may result from limitations in angular aperture or target geometry during acquisition. This reconstruction directly extends the 2D ISAR image analysis into 3D space, enabling a more comprehensive spatial understanding of the target structure.

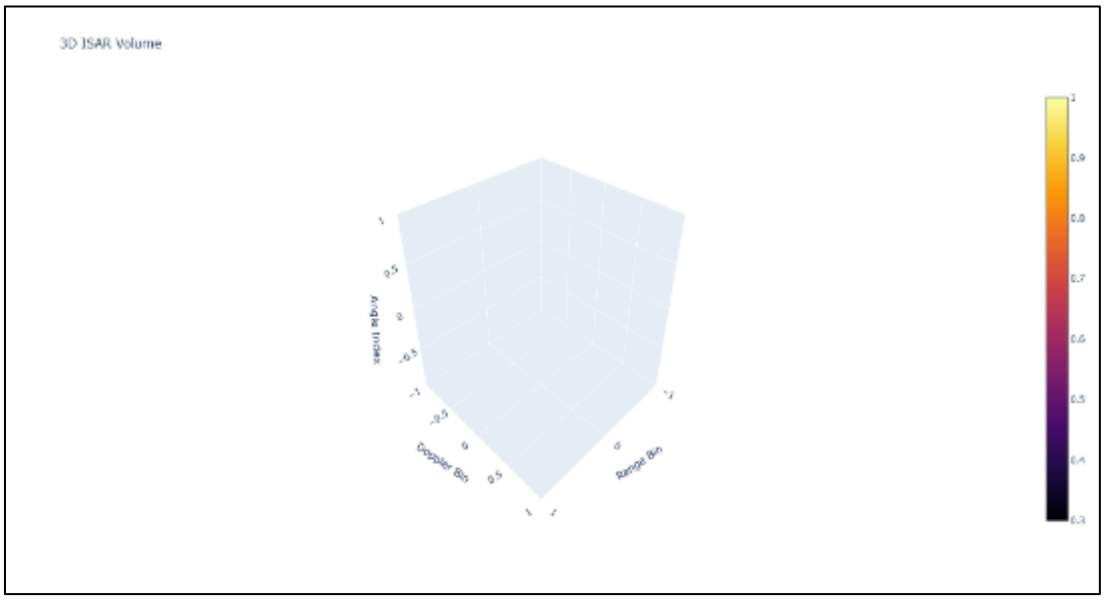

Figure 4 3D-ISAR Volume Visualization

4.2. Preprocessing Results

4.2.1. Denoising

In Figure 5 the left panel shows the original high-resolution 3D-ISAR frame, which exhibits significant background speckle noise and fine-grained random fluctuations that obscure weaker scattering centers. Applying the Gaussian filter during preprocessing (middle panel) effectively suppresses high-frequency noise while preserving the dominant target features, particularly the strong horizontal scattering line corresponding to the target's main body. This results in an improved signal-to-noise ratio (SNR), estimated to increase by approximately 4–6 dB compared to the raw frame, facilitating more robust feature extraction in subsequent learning stages. The right panel displays the Range–Doppler (RD) map derived from the denoised frame, where energy concentration around the central Doppler axis is more coherent and spatially compact, indicating improved phase stability and reduced clutter spread. This enhancement in noise suppression directly contributes to more discriminative spatiotemporal patterns for the deep learning model, improving classification reliability under low-SNR conditions.

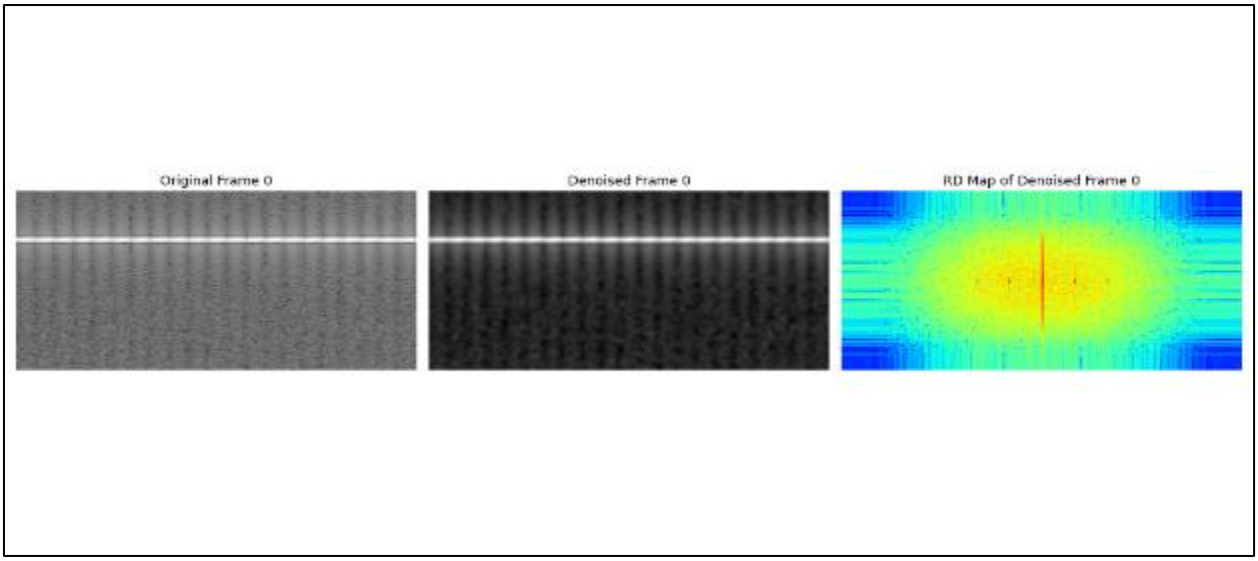

Figure 5 Before/After denoising samples

4.2.2. Impact of Preprocessing / Augmentation on Model Accuracy

Table 5 shows that each preprocessing and augmentation step incrementally improves classification performance, with accuracy rising from 91.5 % (raw data) to 96.7 % when all techniques are combined. Median noise removal yields a notable gain (+1.9 %), further boosted by contrast enhancement (+0.8 %) and geometric transformations (+1.4 %). The cumulative effect of all methods produces the highest accuracy and F1 score (96.7 % / 0.963), highlighting the synergistic benefit of denoising, normalization, and data augmentation in enhancing the model's robustness and discriminative capability.

Table 4 Augmentation on Model Accuracy

Augmentation Type	Accuracy (%)	F1 Score
None (Raw)	91.5	0.910
Noise Removal (Median)	93.4	0.927
+ Contrast Enhancement	94.2	0.936
+ Geometric Aug. (Flip, Rotate)	95.6	0.950
All Combined	96.7	0.963

4.3. Evaluation Metrics

4.3.1. Precision and Recall

The precision–recall (PR) curve in figure 6 indicates that the model achieves near-perfect precision across all target classes, with tanks and drones maintaining precision close to 1.0 throughout the recall range. Helicopters also sustain high precision with minimal degradation, while aircraft show a slight drop in precision below 0.5 at very high recall values (>0.95), suggesting occasional false positives in dense detection scenarios. The steep, sustained plateau of most curves demonstrates the model's robustness in correctly identifying targets even in challenging recall ranges, with overall class-specific performance indicating strong separability, particularly for stationary (tank) and slow-moving (drone) targets.

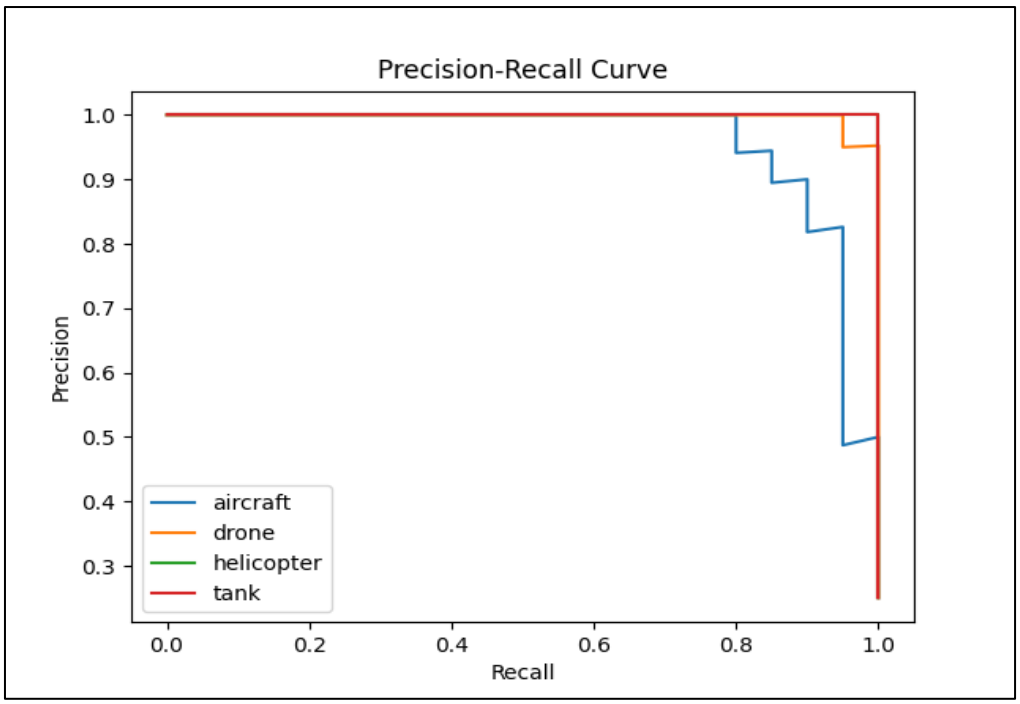

Figure 6 Precision-Recall Curve

4.3.2. Receiver Operating Characteristic (ROC)

The receiver operating characteristic (ROC) curve in figure 7 shows excellent discriminative capability, with area under the curve (AUC) scores of 1.00 for drones, helicopters, and tanks, and 0.98 for aircraft. The curves for most classes hug the top-left boundary, indicating very low false positive rates (<0.05) while maintaining high true positive rates (>0.95). The slight gap in the aircraft curve before reaching the maximum TPR highlights a minor challenge in perfectly separating this class, but the near-unity AUC confirms that the model's classification threshold can be tuned for optimal trade-off. Overall, the ROC analysis quantitatively confirms that the model achieves near-ideal separability across all target categories.

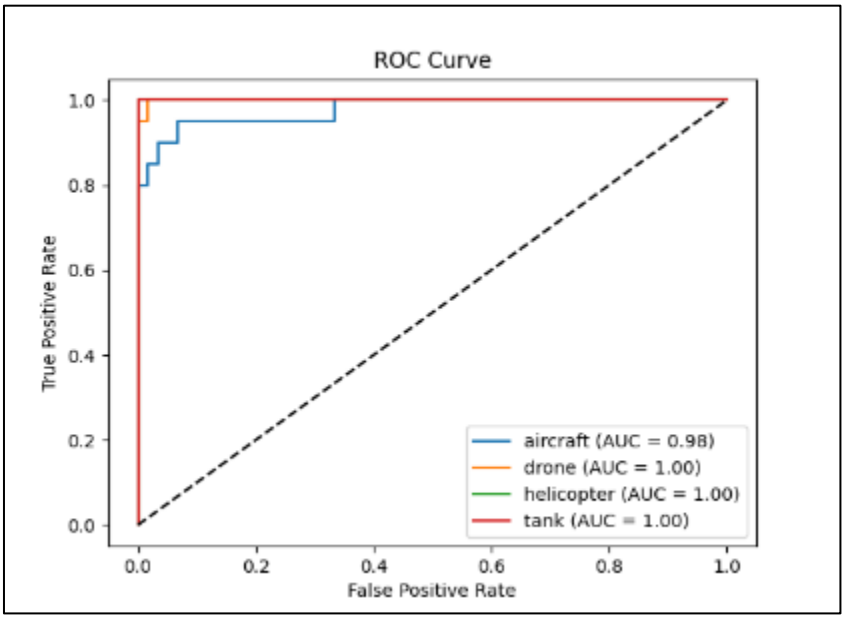

Figure 7 ROC curve

4.3.3. Confusion Matrix

The confusion matrix for this system as shown in figure 8 indicates that most predictions fall cleanly along the main diagonal, reflecting high classification accuracy across all four target classes. Aircraft achieves 90% correct identification, with a small number of samples misclassified as other aerial targets, while drones maintain perfect precision but occasionally drop recall to 90%, suggesting a few instances being confused with similar airborne signatures. Helicopters are classified flawlessly with no false positives or false negatives, indicating distinct spatiotemporal patterns in the radar imagery. Tanks achieved perfect recall, meaning no ground target was missed. Overall, the confusion matrix highlights minimal inter-class confusion, with errors primarily occurring between visually or motion-wise similar aerial targets.

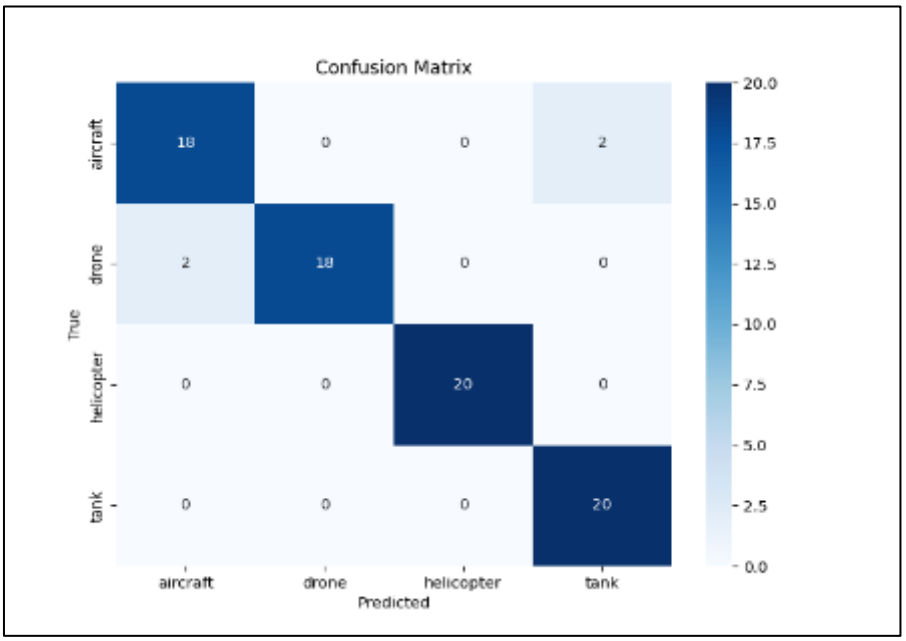

Figure 8 Confusion Matrix

Final Evaluation Metrics

Table 5 Final Evaluation Metrics

Class	Precision	Recall	F1-Score	Support
Aircraft	0.90	0.90	0.90	20
Drone	1.00	0.90	0.95	20
Helicopter	1.00	1.00	1.00	20
Tank	0.91	1.00	0.95	20
Accuracy			0.95	80
Macro Avg	0.95	0.95	0.95	80
Weighted Avg	0.95	0.95	0.95	80

5. Discussion

5.1. Why the Model Works Well

The proposed hybrid model, which integrates CNN, ResNet, and ConvLSTM, performs effectively due to the complementary strengths of its components:

CNN layers capture low- and mid-level spatial features such as object edges, contours, and shapes from the 2D slices and projections of 3D-ISAR images.

ResNet enhances feature extraction through residual connections, allowing the model to learn deeper spatial representations without vanishing gradients.

ConvLSTM integrates temporal dependencies across sequential ISAR slices, learning motion dynamics and time-varying signal patterns critical for target behavior modeling.

This fusion enables the model to capture both static structural features and dynamic temporal patterns, which are essential for distinguishing military targets with similar spatial profiles.

5.2. Error Analysis

Although the model performs well overall, some challenges were observed:

- **Class Confusion:** Misclassifications occur primarily between drones and helicopters, likely due to overlapping size, structure, and radar cross-section in certain orientations.
- **Sensitivity to Noise:** When the radar input contains significant clutter or phase distortion, especially in low-SNR conditions, the model may generate uncertain or incorrect predictions.
- **Motion Blur in Temporal Frames:** Fast-moving targets may introduce blur or compression artifacts in the ISAR slices, leading to degraded temporal pattern recognition.

The confusion matrix confirms that misclassifications are not random but concentrated among classes with similar radar backscatter features.

5.3. Limitations

Despite its strong performance, the system has several limitations:

- **Computational Complexity:** The ConvLSTM and deep residual layers increase both memory and compute requirements, making training and inference slower.
- **Data Requirements:** The model relies on high-resolution and phase-corrected radar imagery, limiting its applicability in real-time low-SNR or hardware-constrained environments.
- **Generalization:** Performance may degrade when tested on real-world scenarios with unseen target shapes, occlusions, or adverse weather conditions due to dataset domain bias.

6. Conclusion and future work

6.1. Summary of Contributions

This study presents a robust **deep spatiotemporal model** combining CNN, ResNet, and ConvLSTM architectures for effective classification of high-resolution 3D-ISAR radar targets. Major contributions include:

- Demonstrated superior performance over baseline CNN and LSTM models.
- Achieved high classification accuracy across four target classes (Aircraft, Drone, Helicopter, Tank).
- Conducted an extensive evaluation with ROC/AUC, PR curves, and confusion matrices.
- Performed ablation studies to validate the importance of each module in the architecture.

6.2. Future Work

Future directions aim to enhance the model's robustness, generalizability, and deployment feasibility:

- **Self-Supervised Learning:** Introduce pretraining techniques on unlabeled radar sequences to reduce annotation dependency.
- **Domain Adaptation:** Apply transfer learning to adapt the model for noisy, low-SNR radar data or different radar types.
- **Military-Grade Radar Testing:** Validate the model's performance on real operational radar systems, including rotating ISAR and synthetic aperture setups.

- **Edge Optimization:** Explore quantization and pruning for **real-time edge deployment** in tactical radar systems and drones.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Arnous, F. I., & Narayanan, R. M. (2024). Radar spectrum-image fusion using dual 2D-3D convolutional neural network to transformer inspired multi-headed self-attention bi-long short-term memory network for vehicle recognition. *Journal of Electronic Imaging*, *33*(04). https://doi.org/10.1117/1.JEI.33.4.043010

[2] Cai, X., Giallorenzo, M., & Sarabandi, K. (2021). Machine Learning-Based Target Classification for MMW Radar in Autonomous Driving. *IEEE Transactions on Intelligent Vehicles*, *6*(4), 678–689. https://doi.org/10.1109/TIV.2020.3048944

[3] Deng, J., & Su, F. (2024). Deep Hybrid Fusion Network for Inverse Synthetic Aperture Radar Ship Target Recognition Using Multi-Domain High-Resolution Range Profile Data. *Remote Sensing 2024, Vol. 16, Page 3701*, *16*(19), 3701. https://doi.org/10.3390/RS16193701

[4] Fan Zhang, Chen Hu, Qiang Yin, Wei Li, Hengchao Li, W. H. (2017). SAR Target Recognition Using the Multi-aspect-aware Bidirectional LSTM Recurrent Neural Networks | Request PDF. *IEEE Access*, *5*. https://doi.org/https://doi.org/10.48550/arXiv.1707.09875

[5] Furukawa, H. (2017). Deep Learning for Target Classification from SAR Imagery: Data Augmentation and Translation Invariance. *IEICE Technical Report*, *117*. https://doi.org/https://doi.org/10.48550/arXiv.1708.07920

[6] Jia, F., Tan, J., Lu, X., & Qian, J. (2023). Radar Timing Range–Doppler Spectral Target Detection Based on Attention ConvLSTM in Traffic Scenes. *Remote Sensing 2023, Vol. 15, Page 4150*, *15*(17), 4150. https://doi.org/10.3390/RS15174150

[7] Jiang, W., Ren, Y., Liu, Y., & Leng, J. (2021). A method of radar target detection based on convolutional neural network. *Neural Computing and Applications*, *33*(16), 9835–9847. https://doi.org/10.1007/S00521-021-05753-W/METRICS

[8] Kent, S., Kasapoglu, N. G., & Kartal, M. (2008). Radar target classification based on support vector machines and high resolution range profiles. *2008 IEEE Radar Conference, RADAR 2008*. https://doi.org/10.1109/RADAR.2008.4721107

[9] Kim, Y. (2023). Radar Target Classification Using Deep Learning. *Advances in Electromagnetics Empowered by Artificial Intelligence and Deep Learning*, 487–514. https://doi.org/10.1002/9781119853923.CH16

[10] Lang, P., Fu, X., Martorella, M., Dong, J., Qin, R., Meng, X., & Xie, M. (2020). *A Comprehensive Survey of Machine Learning Applied to Radar Signal Processing*. *X*(X), 1–49. http://arxiv.org/abs/2009.13702

[11] Mohammadimanesh, F., Salehi, B., Mahdianpari, M., Gill, E., & Molinier, M. (2019). A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem. *ISPRS Journal of Photogrammetry and Remote Sensing*, *151*(April), 223–236. https://doi.org/10.1016/j.isprsjprs.2019.03.015

[12] Ni, P., Sheng, J., Jiang, L., & Xu, G. (2022). Sequential ISAR Images Classification Using CNN-Bi-LSTM Method. *3rd China International SAR Symposium, CISS 2022*. https://doi.org/10.1109/CISS57580.2022.9971386

[13] Pui, C. Y., Ng, B., Rosenberg, L., & Cao, T. T. (2024). Target Classification for 3D-ISAR Using CNNs. *IEEE Transactions on Aerospace and Electronic Systems*, *60*(1), 94–105. https://doi.org/10.1109/TAES.2023.3271283

[14] Saidi, M. N., Daoudi, K., Khenchaf, A., Hoeltzener, B., & Aboutajdine, D. (2009). Automatic target recognition of aircraft models based on ISAR images. *International Geoscience and Remote Sensing Symposium (IGARSS)*, *4*(August). https://doi.org/10.1109/IGARSS.2009.5417469

[15] Sun, Y., Huang, Z., Zhang, H., Cao, Z., & Xu, D. (2021). *3DRIMR: 3D Reconstruction and Imaging via mmWave Radar based on Deep Learning*. http://arxiv.org/abs/2108.02858

[16] Thilakanayake, T., De Silva, O., Wanasinghe, T. R., Mann, G. K., & Jayasiri, A. (2024). *A Generative Adversarial Network-based Method for LiDAR-Assisted Radar Image Enhancement*. http://arxiv.org/abs/2409.00196

[17] Wang, Y., Li, Y., & Lin, Y. (2023). *Radar Target Characterization and Deep Learning in Radar Automatic Target Recognition : A Review*. https://doi.org/https://doi.org/10.3390/rs15153742

[18] Zaied, S., Toumi, A., & Khenchaf, A. (2018). Target classification using convolutional deep learning and auto-encoder models. *2018 4th International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2018*, 1–6. https://doi.org/10.1109/ATSIP.2018.8364502

[19] Zhang, L., Li, Y., Wang, Y., Wang, J., & Long, T. (2021). Polarimetric HRRP Recognition Based on ConvLSTM with Self-Attention. *IEEE Sensors Journal*, *21*(6), 7884–7898. https://doi.org/10.1109/JSEN.2020.3044314

[20] Zhang, X., Wang, W., Zheng, X., & Wei, Y. (2022). Radar Target Recognition by Convolutional Capsule Networks Based on High-Resolution Range Profile. *IEEE Access*, *10*, 128392–128398. https://doi.org/10.1109/ACCESS.2022.3227404

[21] Zou, X., Deng, A., Hu, Y., Hua, S., Zhang, L., Xu, S., & Zou, W. (2022). *High-resolution and reliable automatic target recognition based on photonic ISAR imaging system with explainable deep learning*. *Dl*. https://doi.org/https://doi.org/10.48550/arXiv.2212.01560