IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 29, 2020, accepted October 28, 2020, date of publication November 10, 2020,
date of current version November 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3037062

Deep-Learning Methods for Hand-Gesture Recognition Using Ultra-Wideband Radar

SRUTHY SKARIA[1], (Member, IEEE), AKRAM AL-HOURANI[1], (Senior Member, IEEE), AND ROBIN J. EVANS[2], (Life Fellow, IEEE)

[1]School of Engineering, RMIT University, Melbourne, VIC 3000, Australia
[2]Department of Electrical and Electronic Engineering, The University of Melbourne, Melbourne, VIC 3010, Australia

Corresponding author: Sruthy Skaria (sruthy.skaria@rmit.edu.au)

ABSTRACT Using deep-learning techniques for analyzing radar signatures has opened new possibilities in the field of smart-sensing, especially in the applications of hand-gesture recognition. In this paper, we present a framework, using deep-learning techniques, to classify hand-gesture signatures generated from an ultra-wideband (UWB) impulse radar. We extract the signals of 14 different hand-gestures and represent each signature as a 3-dimensional tensor consisting of range-Doppler frame sequence. These signatures are passed to a convolutional neural network (CNN) to extract the unique features of each gesture, and are then fed to a classifier. We compare 4 different classification architectures to predict the gesture class, namely; (i) fully connected neural network (FCNN), (ii) k-Nearest Neighbours (k-NN), (iii) support vector machine (SVM), (iv) long short term memory (LSTM) network. The shape of the range-Doppler-frame tensor and the parameters of the classifiers are optimized in order to maximize the classification accuracy. The classification results of the proposed architectures show a high level of accuracy above 96 % and a very low confusion probability even between similar gestures.

INDEX TERMS Hand-gesture recognition, deep-learning, radar sensors, radar signal processing, UWB impulse radar.

I. INTRODUCTION

Hand-gesture recognition is gaining significant research interest due to the wide range of envisioned applications. The use of such technology ranges from convenient device control [1], infection prevention in clinical environments [2], to safer and quicker accessibility of features in automotive [3]. The common hand-gesture signal acquisition approaches today are cameras [4], infra-red sensors [5], and ultrasonic sensors [6]. On the other hand, radar sensors are newly emerging due to their superior recognition performance even in adverse lighting conditions and complex background. In addition, low-cost commercial miniature radars are becoming widely available, and are capable of capturing the signature of finer hand movements which can yield high classification accuracy at low processing cost [7].

When using radar sensors for capturing hand-gestures, the type and richness of the gesture signatures depend on the architecture of the radar and on the employed waveform. The common types of waveforms used in miniature radar sensors are; (i) continuous waveform (CW), (ii) pulses, and (iii) frequency modulated continuous waveform (FMCW). Popular CW-based radars are capable of detecting *micro-Doppler* signatures in addition to the main Doppler components; micro-Doppler signatures are the frequency components that occur due to the motion or vibration of the non-rigid parts (fingers, knuckle, wrist) along with the main translational motion of the target (hand) [8]. However, despite the excellent ability to capture Doppler signatures, Doppler radars fail to extract the range information of the targets [9] due to the inherited narrow band nature of the waveform. Simultaneous estimation of range and Doppler is achieved using more advanced waveforms like pulsed wave and FMCW. Therefore, in addition to the Doppler variations, it is possible to distinguish each gesture with the spatial variations of the hand movements. Capturing hand-movements/variations along the radial distance is shown to increase the classification accuracy of the hand-gesture recognition [10].

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

In this study, we utilize a low-power UWB impulse radar which transmits sharp temporal pulses. The advantages of using UWB impulse radar for capturing the range-Doppler signatures compared to their counterparts are mainly; (i) low power consumption, (ii) fine range resolution, and (iii) the ability to detect very close targets [9]. These features make UWB impulse radar an excellent candidate for collecting hand-gestures. In addition, UWB radar has the ability to work reliably in interference-rich environments. The transmitted waveform is extremely short, typically in nanoseconds, hence the signal energy is spread across a very large RF bandwidth providing immunity to the interference [11]. Due to the same reason UWB radars will cause less interference to other devices.

With the collected hand-gesture signatures from the UWB impulse radar, we explore the integration of several machine learning techniques to enhance the hand-gesture classification performance. We present an end-to-end framework for pre-processing the received gesture signals into a sequence of range-Doppler frames forming a 3-dimensional tensor. We utilize 2 different approaches for hand-gesture recognition; the first approach employs a 3D CNN for feature extraction coupled with three different classifiers as the final layer, namely; (i) FCNN, (ii) k-NN, and (iii) SVM. We further present a second approach using 2D CNN along with LSTM to predict the gesture class. The main contributions of this paper are:

- A framework for mapping the raw signal from a UWB impulse radar as a sequence of range-Doppler frames suitable for 3D deep-learning methods.
- Four different CNN architecture models for classifying hand-gesture signatures from a UWB impulse radar.
- Analytic formulation of key controllable parameters to optimize the classification performance.

II. BACKGROUND AND RELATED WORK

The recent developments in consumer radar, motivated numerous researches on integrating radar sensors with machine-learning for hand-gesture recognition. The most common machine-learning approaches for radar-based hand-gesture recognition are CNN, SVM, k-NN, and LSTM. In order to classify hand-gestures using classifiers such as SVM and k-NN, one of the techniques is the manual extraction of hand-gesture features [1], [14], [15] from the range-Doppler or time-frequency (spectrogram) maps. Manual feature extraction requires predefined characteristic features of the gesture signatures, and therefore, the performance of the classifier varies significantly depending on the defined features. Other approaches utilize statistical procedures such as principal component analysis (PCA) for dimension reduction and feature extraction [1], [13]. PCA transforms the input data into a few orthogonal variables (principle components) that represent unique features with reduced dimension.

A common approach in radar hand-gesture recognition is to use CNN, which does not require predefined features, but rather, the network self-learns the features from input signals during the training process [18]. The majority of CNN-based hand-gesture recognition methods extract the signature from either: (i) the changes in Doppler over time [18], or from (ii) a snapshot of the overall range-Doppler fingerprint [19]. Both of these signal types are represented in the form of a 2D matrix (monochromatic image) that is further processed by the CNN. Our previous work [20] utilizes a two-antenna Doppler radar to represent the changes of Doppler over time as 2D spectrogram, along with angle of arrival (AoA) information. The main drawback of 2D image methods is that they lack the 3rd dimension that adds further information to the signature. On the other hand, the representation of range-Doppler maps as a time/frame tensor [22], [23] is shown to increase the richness of the signatures, thus, leading to a better description of the hand-gestures. For the range-Doppler-frame tensor, suitable features can be extracted using a 3D CNN, a process that is followed in [23], [26], [27]. Another recent research demonstrates that integrating CNN and LSTM tends to increase the classification performance of the signatures that vary in time and space [22], [23]. LSTM networks are recurrent neural networks with feedback connections, which makes them suitable for time sequence analysis [21]. We describe in Table. 1 some of the approaches found in the literature for hand-gesture recognition using radar sensors, including the utilized machine-learning algorithms, the type of radar waveform, and the set of gesture signatures used.

To the best of the authors knowledge, this paper is the first to provide an end-to-end framework for hand-gesture recognition using the range-Doppler-frame tensor from a UWB impulse radar in integration with multiple deep-learning methods.

III. SYSTEM DESCRIPTION

In this proposed framework, we utilize a UWB impulse radar with a single-output-single-input configuration, i.e. one antenna for transmitting and one for receiving. The reflected electromagnetic signal is captured by the receiving antenna, then sampled in the RF domain and digitally downconverted to baseband. The output of the radar module (the baseband signal) is passed to the computer as two vectors; (i) in-phase component (I) and (ii) quadrature component (Q). The received signal of each gesture is processed to form a 3-dimensional tensor of range-Doppler-frame, as shows in Fig. 1. This tensor represents the pattern generated by a hand-gesture as a sequence of frames each consisting of a temporal snapshot of range-Doppler image.

Fig. 2 shows a functional block diagram of the classification framework. In the proposed framework we present 4 different classifiers and compare their performance; (i) 3D CNN for feature extraction with FCNN for classification, (ii) 3D CNN feature extractor and k-NN classifier, (iii) 3D CNN feature extractor and SVM classfier, (iv) 2D CNN feature extractor and LSTM classifier. The range-Doppler-frame tensor is passed to a CNN to extract the features representing

TABLE 1. Literature review.

Work summary	Ref.	Approach	Waveform	Gesture signature
Utilizes manual extraction of predefined features from the gesture signals received using a UWB impulse radar. It employs k-NN to classify the gestures using the features.	[12]	k-NN	UWB impulse radar	Time domain signal
PCA is used to extract the features from the gesture signals received using an UWB impulse radar. Then, an FCNN is utilized to train and classify the hand-gestures.	[13]	PCA - FCNN	UWB impulse radar	Time domain signal
Compares the performance of PCA based classification and classification based on manual feature extraction from the gesture signals.	[1]	PCA - k-NN	Doppler radar	Spectrogram
Propose a technique to extract feature vector from the envelopes of the micro-Doppler signatures. Then apply a k-NN classifier to classify the gestures using the feature vector.	[14]	k-NN	Doppler radar	Spectrogram
Predefined features are extracted from the spectrogram of the received gesture signals, which are then utilized to classify the gestures employing an SVM classifier.	[15]	SVM	Doppler radar	Spectrogram
The work utilizes the time-frequency analysis to represent the gesture signatures received using an FMCW radar and employs a single channel CNN to classify nine gestures.	[16]	2D CNN	FMCW	Spectrogram
The reflected signal from the hands are received using a UWB impulse radar and is passed to a 1D CNN for classification. Further, the work compares the classification performance of the CNN classifier and SVM classifier.	[17]	1D CNN	UWB impulse radar	Time domain signal
Hand-gesture signals are captured using a Doppler radar are pre-processed using short time discrete Fourier transform (STDFT) to obtain the spectrograms, which are fed to an optimized single channel CNN network.	[18]	2D CNN	Doppler radar	Spectrogram
The spectrograms of the received hand-gesture signals from four receiver channels of an FMCW radar is mapped to the input of CNN, where data fusion is carried out at an adjustable position.	[19]	2D CNN	FMCW	Spectogram
In our previous work the receiving antennas of a continuous-wave Doppler radar capable of producing the in-phase and quadrature components of the beat signals is utilized to map into the input channels of a CNN as two spectrograms and an AoA matrix.	[20]	2D CNN	Doppler radar	Spectogram & AoA matrix
Range-Doppler maps (RDM) from four receiver channels of an FMCW radar are processed to obtain sequential feature vectors called projected RDM, which is mapped to an optimized LSTM network for classification.	[21]	1D LSTM	FMCW	RDM
The authors propose an end-to-end classification network with 2D CNN and LSTM. The gesture features are extracted from range-Doppler images using the CNN and classify them using the subsequent LSTM layer.	[22]	2D CNN - LSTM	FMCW	Range-Doppler-frames
Proposes a 3D CNN architecture to learn the embedding model using distance-based triplet-loss similarity metric and classify the gestures using k-NN by utilizing the distance metric.	[23]	3D CNN - k-NN	FMCW	Range-Doppler-frames
Proposes an algorithm that transforms 2D image data of the trajectory of hands into trigonometric ratios and plots them against the time axis to obtain unique images of the gestures and utilize a CNN for classification.	[24]	2D CNN	Multiple UWB radars	Trajectory image
Utilizes three UWB radars to extract hand's mid-air trajectory and then employee CNN for the classification of hand-gestures. The proposed method uses digits written in the air as the hand-gestures/movements.	[25]	2D CNN	Multiple UWB radars	Trajectory image

different hand-gestures. Using the extracted features we train the 4 different classifiers, FCNN, k-NN, SVM, and LSTM which are illustrated in Fig. 3 to predict the classes of hand-gestures. The following subsections provide a detailed description of each component in the proposed framework, from hand-gesture collection to gesture classification.

IV. HAND-GESTURE COLLECTION AND PRE-PROCESSING

In order to collect the hand-gesture signatures we utilize a Xethru X4M03 UWB impulse radar module from Novelda [28] which is shown in Fig. 4. The parameters of the radar are tuned to fit the requirements of the hand-gesture recognition application which are described in Table. 3. The selected gestures are the typical movements of hands. We utilize a total of 14 hand-gestures for the study, which are depicted in Fig. 5. The gestures are performed using the right hand. The description of the selected 14 gestures is given in Table. 2. In order to introduce variations in the collected gestures, we perform the gesture collection in arbitrary radar orientations with respect to the surrounding room environment with randomized speeds and distances. These variations would increase the richness of the data set, allowing for a better classification performance.

A. UWB IMPULSE RADAR

A UWB impulse radar [29] transmits a sequence of short pulses (in our case Gaussian-shaped) having a duration/width T_p in the order of nanoseconds (ns). The main difference

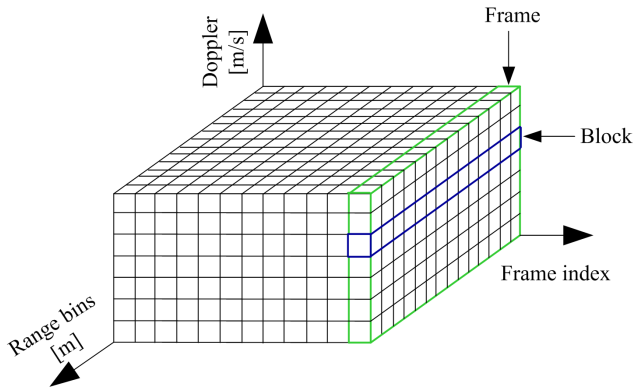

FIGURE 1. An illustration of 3-dimensional range-Doppler-frame tensor.

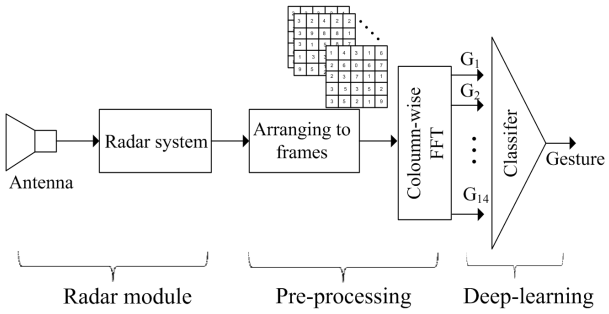

FIGURE 2. The proposed framework for mapping the signatures generated by a UWB impulse radar into deep-learning classifiers.

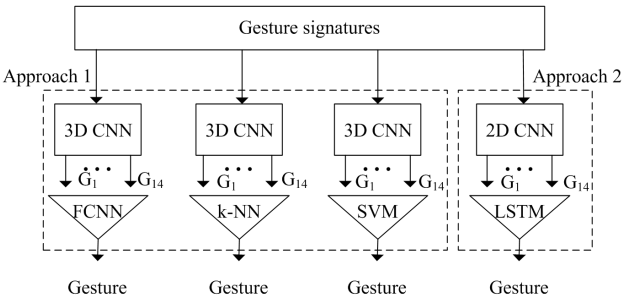

FIGURE 3. The utilized approaches for feature extraction and classification of gesture signatures from the range-Doppler-frame tensor.

between a UWB impulse radar and a standard pulse radar is in the utilized pulse width. A UWB impulse radar transmits short pulses of pulse width comparable with the period of the carrier waveform, whereas the typical pulse radar utilizes pulses with pulse width larger than many periods of the carrier waveform. The short pulses of the UWB impulse radar provides a wide bandwidth, which in turn yields a high range resolution, given as,

$$\Delta R = \frac{c}{2B}, \tag{1}$$

where B is the bandwidth and c is the propagation speed of light.

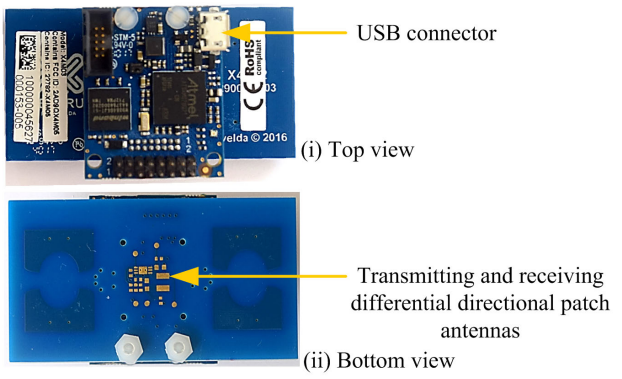

FIGURE 4. The UWB impulse radar module used for data collection in the experiment [28].

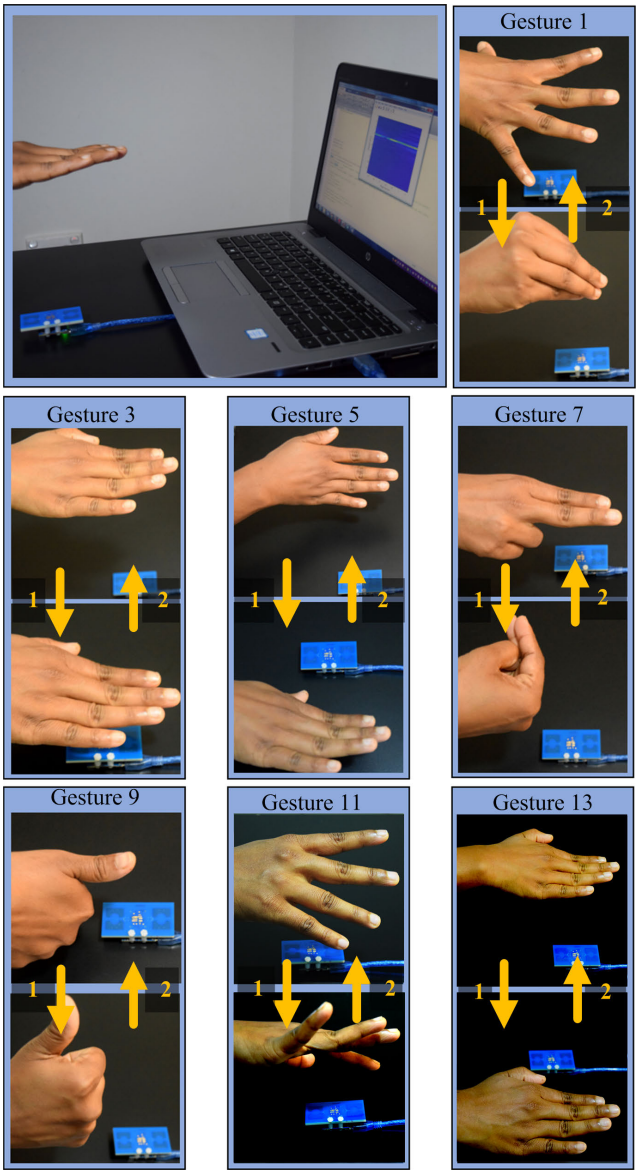

FIGURE 5. The pictures of the performed hand-gestures in this paper.

In order to get better insights on the range-Doppler processing, we present some details on how the signal is received

TABLE 2. Employed gestures.

Gesture (1) Single blinking: Opening then closing of fingers together	Gesture (2) Double blinking
Gesture (3) Single to-and-fro: Moving open hand towards and away from the radar	Gesture (4) Double to-and-fro
Gesture (5) Single round: Rotating hand in an anti-clockwise direction	Gesture (6) Double round
Gesture (7) Single swiping: Swiping using two fingers	Gesture (8) Double swiping
Gesture (9) Single thumbs up: Moving hand towards and away with thumbs-up	Gesture (10) Double thumbs-up
Gesture (11) Single waving: Fingers waving while the palm stays still	Gesture (12) Double waving
Gesture (13) Single sliding: Sliding the hands from left to right	Gesture (14) Double sliding

and processed inside the utilized radar module before being collected for further processing. Fig. 6 illustrates the architecture of the utilized UWB impulse radar module. The utilized radar module employs a direct-RF synthesizer [2] to generate the Gaussian pulses with an analytic signal form,

$$g(t) = A \exp\left(j2\pi f_o t\right) \exp\left(\frac{-t}{T_p}\right)^2, \tag{2}$$

where the carrier frequency f_c is modulated with a baseband Gaussian pulse $A \exp\left(\frac{-t}{T_p}\right)^2$, having an amplitude A.

The reflected impulse from the hand is received at the receiver given by,

$$r(t) = \exp\left(j2\pi f_o\left(t-\tau\right)\right) \exp\left(\frac{-\left(t-\tau\right)^2}{T_p^2}\right), \tag{3}$$

where, $\tau = \frac{2R}{c} + \frac{2vt}{c}$ is the time delay, R and v are the range and radial velocity of the target respectively. The signal is reconstructed at the receiving side of the radar module using the swept-threshold sampling method [30]. The $r(t)$ is sampled with a high sampling rate f_s (In our application, for practical reasons, the high-rate sampling is implemented by employing 12 parallel samplers of sampling frequency $\frac{f_s}{12}$, each sampling with a slight delay equivalently giving a sampling rate f_s [28]). The received pulse samples are thresholded to V using a comparator (output of the comparator is 1 and 0, where 1 for signal above the threshold and 0 otherwise). After certain number of pulses, the threshold voltage V is stepped using a digital-to-analog converter (DAC). As a result, V is swept between $[V_{min}, V_{max}]$. The comparator output is also sampled at a sampling rate f_s and are distributed across N counters.

The sampled bits are summed, across each counter to incrementally build the multi-bit block which gives a cumulative distribution function of the received signal. This digitally reconstructed RF signal block is

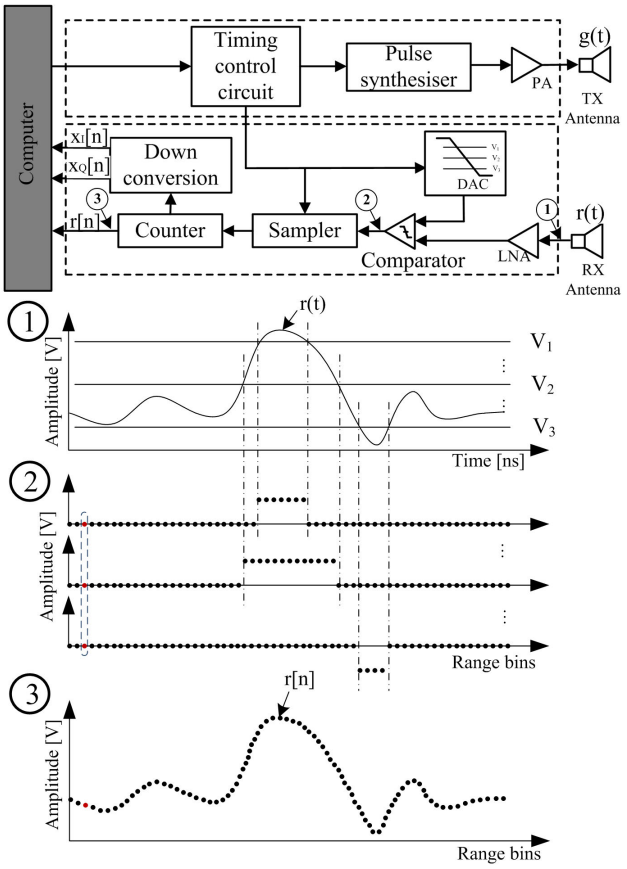

FIGURE 6. The block diagram of the utilized UWB impulse radar module.

given by,

$$r[n] = \exp\left(j2\pi f_o\left(t_n-\tau_n\right)\right) \exp\left(\frac{-\left(t_n-\tau_n\right)^2}{T_p^2}\right), \tag{4}$$

where, n is the discrete time index. A simulation of the generated Gaussian pulse is given in Fig. 7. The number of counters N is the same as the number of range bins, which gives the maximum range,

$$R_{max} = \frac{t_{max} \times c}{2} = \frac{N t_s \times c}{2} = \frac{\frac{N}{f_s} \times c}{2}, \tag{5}$$

where $t_s = \frac{1}{f_s}$ is the sampling period. Therefore, a block represents the strength of reflection located in each range bin. Thus, several such blocks are collected as the output of the radar module. We can read out this RF data directly or enable on-chip digital down conversion to read the baseband analytic signal where, we utilize the on-chip down conversion to obtain the baseband signal. After down conversion we get a complex block, as the in-phase (I) component, $x_I[n] = \mathrm{re}\{x[n]\}$ and the quadrature (Q) component, $x_Q[n] = \mathrm{im}\{x[n]\}$ of the baseband signal

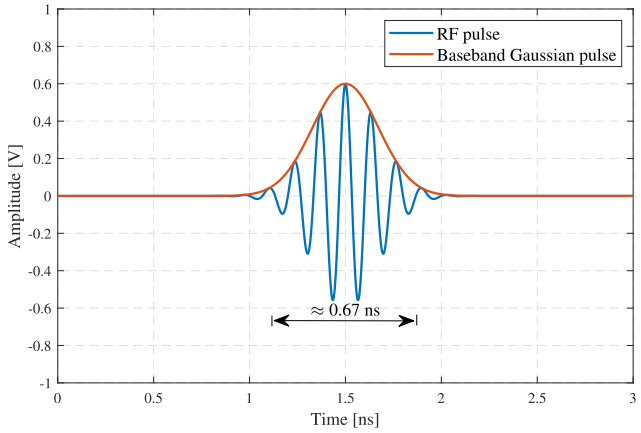

FIGURE 7. Simulation of digitally generated Gaussian pulse with the reconstructed pulse using swept threshold sampling.

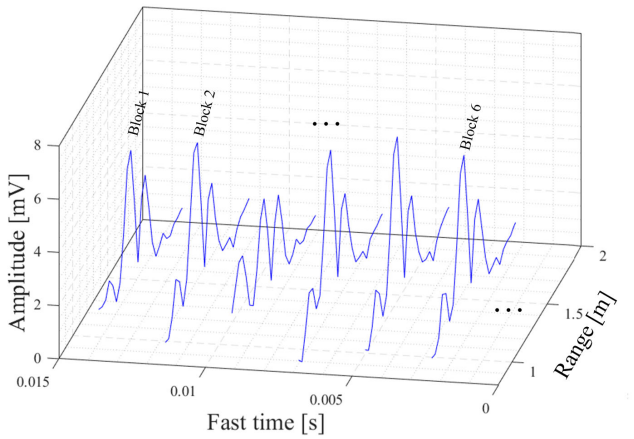

FIGURE 8. An example of a set of blocks as received from the radar module.

given by,

$$x[n] = \exp\left(j2\pi f_o n_\tau\right) \exp\left(\frac{-(n-n_\tau)^2}{T_p^2}\right). \tag{6}$$

Fig. 8 illustrates a down converted block of the received signal. The timing control unit of the module digitally controls the entire process. The summary of parameters employed in the radar sensor is given in Table 3.

B. RANGE-DOPPLER FRAMES

As explained in section IV-A, at a given time instance the radar interprets the scene as a block of range bins. The block rate of a UWB impulse radar is given by, $f_{\text{block}} = \frac{f_{\text{PRF}}}{K}$, where f_{PRF} is the pulse repetition rate, and K is the number of pulses per block. We arrange each of the M blocks into a single frame, such that the changes in the stacked blocks are used for extracting Doppler information using the fast Fourier transform (FFT). The resulting Doppler resolution is given by,

$$\Delta_f = \frac{f_{\text{block}}}{M}, \tag{7}$$

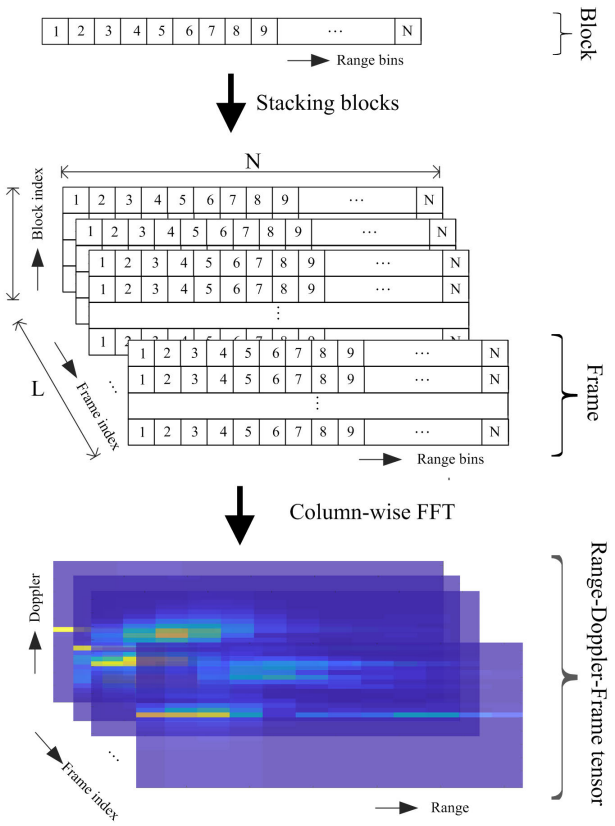

FIGURE 9. Illustration of the range-Doppler-frame tensor and the main steps to formulate it, starting from the raw signal captured by the UWB impulse radar.

However, it can be shown that the Doppler frequency is given by $\frac{2vf_o}{c}$, thus $f_{\text{block}} = \frac{2M\Delta_v f_o}{c}$, where Δ_v is the velocity resolution. The block is related to the frame rate as $f_{\text{frame}} = \frac{f_{\text{block}}}{M}$. Accordingly, there is a trade-off between the frame rate and velocity resolution as follows,

$$\frac{f_{\text{frame}}}{\Delta_v} = \frac{2f_o}{c}. \tag{8}$$

Furthermore, in-order to capture the changes in range-Doppler signature over time we stack each L frames into a 3D matrix (tensor) as indicated in Fig. 9. Therefore, we optimize f_{frame} and Δ_v to obtain the maximum classification accuracy. The range-Doppler-frame tensor has a dimension $N \times M \times L$. However, the dimension of the tensor is slightly different as we cut some unwanted values in all the three dimensions, which we explain in the following paragraph.

We define the length of the gesture in time as *total sample time* which is kept constant for all the 14 gestures. The total sample time for each gesture is set as 3 seconds. In-order to collect only the useful signal (reflected signal once the hand is present), we applied a suitable threshold to crop the received signal. Since the hand-gestures are taken in the range 0.5 to 1 m from the radar module, we took only 15 range bins starting from 0.4 to 1.3 m. Thereafter, we compute column-wise FFT of each frame to get the range-Doppler-

TABLE 3. Radar parameters.

Parameter	Symbol	Value
Center frequency	f_o	7.29 GHz
Bandwidth	B	1.5 GHz
Peak pulse output power	–	−0.7dBm
Antenna beam width azimuth	–	65°
Antenna beam width elevation	–	65°
Pulse repetition frequency	f_{PRF}	15.1875 MHz
Sampling rate	f_s	23.382 GHz
Block rate	f_{block}	440 Hz
Gaussian pulses width	T_p	0.67 ns
Range resolution	ΔR	0.0514 m
Maximum range	$R_{\max}$	9.9 m
Velocity resolution (Network I)	Δ_v	0.23 m/s
Velocity resolution (Network II)	Δ_v	0.08 m/s
Velocity resolution (Network III)	Δ_v	0.15 m/s
Velocity resolution (Network IV)	Δ_v	0.11 m/s

frame tensor. In-order to capture the gestures we use a block rate of $f_{\mathrm{block}} = 440$ Hz, which gives a maximum Doppler frequency of $\frac{f_{\mathrm{block}}}{2} = \pm 220$ Hz. This range is much larger than the observed Doppler frequencies of the performed hand-gestures which are within $\Delta_{f_{\max}} = \pm 120$ Hz. The frequency range is restricted to retain only the significant signal components in the range within ±120 kHz. Therefore, the reduction of points in all the three dimensions significantly reduces the computational load on the classification networks.

V. CLASSIFICATION ARCHITECTURES

A. 3D CNN ARCHITECTURE

Recent developments in 3D CNN have proven to be effective for the classification of volumetric data such as video [31], computed tomography images [32], magnetic resonance imaging [33], ultrasound imaging [34] etc. 3D CNN has the same principles as the 2D version, it is composed of a series of basic structures repeated multiple times. The basic structure is primarily composed of: (i) a *convolutional layer* intended for feature extraction, (ii) activation function for a non-linear transformation of the inputs, and (iii) a pooling layer to reduce the dimension and noise of the input [35]. The difference between 3D and 2D CNN is that in 3D CNN, mathematical operations are done using 3D matrices (tensors) [36], this naturally requires higher processing and memory capacity. In our application, the 3D CNN extracts the temporal variation along with the range-Doppler features from the 3D data tensor as explained in the previous section.

In a typical CNN architecture, once the features are extracted, the classification is performed by a fully connected layer (conventionally an FCNN [35]). Whereas in this work, along with the FCNN we explore the performance of two more classifiers: k-NN and SVM, when integrated with the 3D CNN. Following we describe how these classifiers are integrated,

TABLE 4. Summary of 3D CNN - FCNN.

Layer	Shape	Parameters
Input shape	25x15x29	-
3D convolutional layer	7, 5x5x5	ReLU
Max-pooling layer	1x1x1	
3D convolutional layer	11, 5x5x5	ReLU
Flatten layer	-	-
	FCNN classifier	
Fully connected layer	128	ReLU
Final fully connected layer	14	Softmax

TABLE 5. Summary of 3D CNN - *k*-NN.

Layer	Shape	Parameters
Input shape	67x15x10	-
3D convolutional layer	20, 3x3x3	ReLU
Max-pooling layer	1x1x1	-
Flatten layer	-	-
Fully connected layer	1000	ReLU
	k-NN classifier	
Neighbours	1	-
Weights	-	Uniform
Metric	-	Minkowski

1) 3D CNN - FCNN (Network I)

FCNN is a feed forward neural network with fully connected multi-layer perceptron (MLP), having all the inputs from one layer connected to the input of the next layer. The summary of the employed architecture is shown in Table 4, we also reference to this architecture as *Network I*.

2) 3D CNN - K-NN (Network II)

k-NN classifies an input by identifying the class based on the proximity of the outcoming pattern (from the 3D CNN) to the pre-trained classes. In particular, it takes the majority votes from its k nearest neighbours and assign the outcome class based on the dominant number of neighbours. We optimize the k value to obtain the maximum classification accuracy which is for $k = 1$, which means the input is simply assigned to the class of the nearest neighbor. The summary of the employed 3D CNN - k-NN architecture is given in Table 5, we also reference to this architecture as *Network-II*.

3) 3D CNN - SVM (Network III)

SVM classifiers are supervised learning models that construct a set of hyper-planes in a higher-dimensional space to separate each class [37]. It utilizes support vectors, which are the data points that determines the hyper-planes to separate the classes. The summary of the employed 3D CNN - SVM architecture is given in Table 6, we also reference to this architecture as *Network-III*.

B. 2D CNN - LSTM (Network IV)

LSTM networks comes under the recurrent neural network (RNN) group, which has the ability to analyse time-series inputs [38]. The network consists of cell-state/memory, which makes it possible to store data from

TABLE 6. Summary of 3D CNN - SVM.

Layer	Shape	Parameters
Input shape	73x15x9	-
3D convolutional layer	5, 5x5x5	ReLU
Max-pooling layer	1x1x1	-
Dropout layer	0.5	-
3D convolutional layer	5, 5x5x5	ReLU
Flatten layer	-	-
	SVM classifier	
Kernal		Radial basis function
Regularization	3	-
Gamma		Scale

TABLE 7. Summary of 2D CNN - LSTM.

Layer	Shape	Parameters
Input shape	14x49x15	-
2D convolutional layer	7, 7x7	ReLU
Flatten layer	-	-
Fully connected layer	500	ReLU
Time Distributed layer	-	-
LSTM	14	-
Final fully connected layer	14	Softmax

previous state/time. Cell-states carry relevant information throughout the processing of the time sequence, thereby, identifying and extracting the temporal relation within the sequence [39]. Along with the cell-states, LSTM consists of 3 different gates, which are basically neural networks that control the weights and outputs at each state [40]. These three gates are as follows:

- Forget gate: The information from the previous state and current input are utilized to decide on which cell state/memory to be deleted/forget and which ones to be kept.
- Input gate: It decides on which values of the cell state need to be modified/updated.
- Output gate: Output gate utilizes the updated cell state and the current inputs to compute the weights of the current state.

These gates control the weights and outputs of the cell at each time step establishing temporal relations. Once the temporal features are extracted, a classifier, usually a FCNN is used for classification of the inputs. Since the hand-gestures are represented as a temporal sequence of range-Doppler-frame, at each time step a 2D CNN is utilized to extract features from the corresponding range-Doppler frame. Thus, the LSTM network establishes the temporal relation between the features of the range-Doppler-frame tensor. The end-to-end network is trained using the back-propagation algorithm. A summary of the network architecture is provided in Table 7, we also reference to this architecture as *Network-IV*.

VI. EXPERIMENTATION AND VALIDATION

In-order to perform experimental verification of the proposed framework, we collect 250 samples of each of the 14 gestures. The samples are divided into two groups, (i) the first contains 80% of samples and is used for training the network with 5-fold cross validation, while (ii) second group with the remaining 20% of the samples are only used for testing the networks without being in the training process. This grouping will better allow to understand the performance under realistic use scenarios where recognizing *unseen* data is required.

We first optimize the dimensions of the input to obtain the best classification accuracy and using the selected input parameters of the networks such as, number of layers, kernel size, and number of features in the convolutional layers, number of FCNN, k in k-NN, kernel functions in SVM, are optimized to obtain the architecture which has the maximum classification accuracy. The sensitivity of the classification networks to the other parameters are observed as negligible. We use Python programming language, especially Keras library to build and train the networks. Fig. 10 shows the comparison of the classification accuracy obtained for each classifier for different values of M ranging from 20 to 100 blocks per frame. The value of M determines the input size of the range-Doppler-frame tensor. The range of M we tested is limited to $20 - 110$ is because, beyond this range, the size of the input will fall below the minimum size requirements of the filer sizes of the CNN network [20]. The experimental results of each classification network are detailed below.

- Network I: The network gives a classification accuracy of 93.33 % for 14 gestures. The highest classification accuracy is obtained for an input size of the tensor, $25 \times 15 \times 29$ with $M = 40$ blocks per frame, where the dimension represents Doppler, range, and frame respectively. We select to use 10 epochs for the training with a batch size of 20 samples.
- Network II: A classification accuracy of 92.02 % is noted with the replacement of the FCNN with the k-NN classifier. Maximum classification performance is achieved for an input size $67 \times 15 \times 10$ with $M = 110$ blocks per frame. The k-NN classifier performance is slightly less than the conventional 3D CNN with FCNN. The reason can be due to the similarities of the hand-gestures that could make the distance metric sensitive.
- Network III: The classification network using 3D CNN-SVM shows higher classification performance compared to the previous methods. An overall classification accuracy of 94.08% for 14 gestures. Integration of SVM with CNN is a novel approach for hand-gesture recognition using radar sensors. Maximum classification performance is achieved with an input size $73 \times 15 \times 9$ with $M = 60$ blocks per frame.
- Network IV: The network shows the highest classification performance in classifying 14 gestures with an accuracy of 96.15 %. The network gives maximum classification performance for $M = 80$ blocks per frame with the input size of $14 \times 19 \times 15$ with dimensions time, Doppler, and range respectively. We use 20 epochs with a batch size 20 samples for the training of the network.

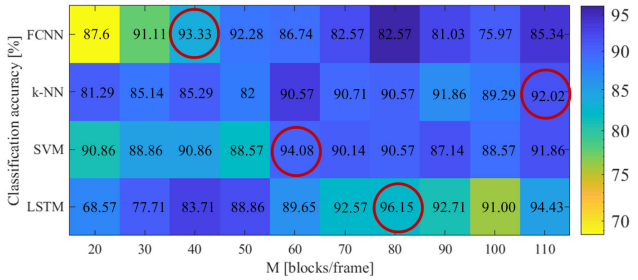

FIGURE 10. Average classification accuracy of the four different classifiers with varying block rate.

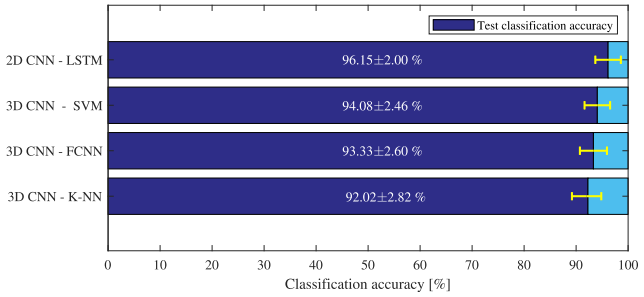

FIGURE 11. Comparison of the classification performance of the four classifiers based on average hand-gesture classification accuracy.

The testing classification accuracy of the Network IV outperforms the Network I by 3 %, outperforms the Network II by 4 % and, outperforms Network III by 2 % as indicated in Fig. 11. The computation time for training Network I is approximately 30 minutes, for Network II is around 12 minutes, for Network III is approximately 27 minutes, and for Network IV is around 14 minutes using a typical laptop with an Intel Core i5 processor. However, there is no significant difference in the recognition time (prediction time during the testing which is in milliseconds) of the gesture by the networks. The computational complexity in terms of memory usage and trainable parameters is highest for the Network II, the second for the Network III network, third for the Network I and lowest computational cost is for the Network IV, which gives the highest classification performance.

Table 8 shows the average confusion matrix between gestures obtained from the Network IV based on the training results. From the confusion matrix it can be observed that the highest confusion appears among gestures 7 and 8 and 11 and 12. The reason being the limited radial movements for these gestures which limits the Doppler signatures which in turn affects the classification of these gestures. Despite that, the overall performance of the Network IV is promising in classifying similar gestures.

Given the sample size of $n = 250$, and selecting confidence interval as $p = 90\%$, the error bounds ϵ are calculated [20] based on the obtained success probability estimate $\hat{\beta}$ using,

$$\epsilon = z_p \sqrt{\frac{\hat{\beta}(1-\hat{\beta})}{n}},$$

TABLE 8. Confusion matrix.

Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	**90.57**	9.75	0	0	0.96	0	0	0	0	1.78	0	0	0	0
2	0	**100**	0	0	0	0	0	0	0	0	0	1	0	0
3	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0
4	0	0	6	**94.54**	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0
6	0	0	0	3.63	0	**95.83**	0	0	0	0	0	0	0	0
7	0	4.87	0	0	0	0	**96**	0	0	0	0	0	0	0
8	0	0	0	0	0	0	12	**88.23**	0	0	0	0	0	0
9	1.8	4.8	0	0	0	0	0	0	**94.73**	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0
11	0	0	0	0	0	0	2	0	0	0	**94.64**	3.44	0	0
12	0	0	0	0	0	0	0	1.96	0	0	10.71	**87.93**	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0
14	0	0	0	0	0	0	0	0	0	0	0	0	5	**95.34**

where z_p denotes the inverse-CDF of a standard normal distribution (quantile function) calculated at the probability $1 - \frac{1-p}{2}$ with p confidence. The error bar is also shown in Fig. 11. Considering the presented results of the Network IV with a training accuracy of 98.40 % and testing accuracy 96.15 % with 14 gestures, the proposed architecture in this paper outperforms our previous work [20] that was based on two-antenna Doppler radar with CNN classifier (having an accuracy of 95.5%), it also outperforms recent CNN-LSTM approach in the literature [23] utilizing a FMCW radar (having an accuracy of 88.1% with 6 gestures). These enhancements are primarily referred to the higher range resolution offered by low-cost UWB radar compare to their counterparts such as FMCW at the same cost range. The higher range resolution is due to the high bandwidth of the UWB radars that in-turn gives richer gesture signatures. Also, the accuracy improvements are caused by the optimization of both the tensor and the network parameters.

In order to observe the behaviour of the Network IV on multiple volunteers we collected gestures from two volunteers; one female volunteer and one male volunteer so that we increase the variations in the received signals. With the proposed architecture we trained the network using separate data sets from each volunteer. The classification accuracy obtained for individual samples from volunteer 1 is 96.15% and with volunteer 2 is 91.5%. The proposed method is showing favourable results for practical applications.

VII. CONCLUSION

This paper investigated the feasibility of using four different classifiers for recognizing hand-gestures from a UWB impulse radar. It presented a novel framework for mapping the output of the UWB impulse radar into a sequence of range-Doppler frames that are fed to CNN-based classifiers. The classification accuracies for the 14 hand-gestures were quite high (93.33 %, 92.02 %, 94.08 % and 96.15 %) for the four different classifiers: (i) 3D CNN - FCNN (ii) 3D CNN - k-NN (iii) 3D CNN - SVM (iv) 2D CNN - LSTM. This indicates the considerable promise for utilizing UWB radar in practical hand-gesture applications.

ACKNOWLEDGMENT

The authors would like to thank D. Huang (RMIT University, Melbourne, VIC, Australia) for providing gesture samples to verify the performance of the proposed algorithm.

REFERENCES

[1] Q. Wan, Y. Li, C. Li, and R. Pal, "Gesture recognition for smart home applications using portable radar sensors," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 6414–6417.

[2] N. Andersen, K. Granhaug, J. A. Michaelsen, S. Bagga, H. A. Hjortland, M. R. Knutsen, T. S. Lande, and D. T. Wisland, "A 118-mW pulse-based radar SoC in 55-nm CMOS for non-contact human vital signs detection," *IEEE J. Solid-State Circuits*, vol. 52, no. 12, pp. 3421–3433, Dec. 2017.

[3] S. Skaria, A. Al-Hourani, R. J. Evans, K. Sithamparanathan, and U. Parampalli, "Interference mitigation in automotive radars using pseudo-random cyclic orthogonal sequences," *Sensors*, vol. 19, no. 20, p. 4459, Oct. 2019.

[4] B. Ionescu, V. Suse, C. Gadea, B. Solomon, D. Ionescu, S. Islam, and M. Cordea, "Using a NIR camera for car gesture control," *IEEE Latin Amer. Trans.*, vol. 12, no. 3, pp. 520–523, May 2014.

[5] F. Erden and A. E. Cetin, "Hand gesture based remote control system using infrared sensors and a camera," *IEEE Trans. Consum. Electron.*, vol. 60, no. 4, pp. 675–680, Nov. 2014.

[6] Y. Sang, L. Shi, and Y. Liu, "Micro hand gesture recognition system using ultrasonic active sensing," *IEEE Access*, vol. 6, pp. 49339–49347, 2018.

[7] A. Al-Hourani, R. Evans, P. M.Farrell, B. Moran, S. Kandeepan, Skafidas, and U. Parampalli, "Millimeter-wave integrated radar systems and techniques," in *Academic Press Library in Signal Processing* (Array, Radar and Communications Engineering), vol. 7, S. Theodoridis and R. Chellappa, Eds. Amsterdam, The Netherlands: Elsevier, 2016, ch. 7.

[8] D. Tahmoush, "Review of micro-Doppler signatures," *IET Radar, Sonar Navigat.*, vol. 9, no. 9, pp. 1140–1146, Dec. 2015.

[9] X. Li, Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote Sens.*, vol. 11, no. 9, p. 1068, May 2019.

[10] O. R. Fogle and B. D. Rigling, "Micro-range/micro-Doppler decomposition of human radar signatures," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 48, no. 4, pp. 3058–3072, Oct. 2012.

[11] X. Wang, A. Dinh, and D. Teng, "Radar sensing using ultra wideband—Design and implementation," in *Ultra Wideband–Current Status and Future Trends*, Rijeka, Croatia: InTech, Oct. 2012.

[12] F. Khan, S. Leem, and S. Cho, "Hand-based gesture recognition for vehicular applications using IR-UWB radar," *Sensors*, vol. 17, no. 4, p. 833, Apr. 2017.

[13] J. Park and S. H. Cho, "IR-UWB radar sensor for human gesture recognition by using machine learning," in *Proc. IEEE 18th Int. Conf. High Perform. Comput. Commun., IEEE 14th Int. Conf. Smart City; IEEE 2nd Int. Conf. Data Sci. Syst. (HPCC/SmartCity/DSS)*, Dec. 2016, pp. 1246–1249.

[14] M. G. Amin, Z. Zeng, and T. Shan, "Hand gesture recognition based on radar micro-Doppler signature envelopes," in *Proc. IEEE Radar Conf. (RadarConf)*, Apr. 2019, pp. 1–6.

[15] S. Zhang, G. Li, M. Ritchie, F. Fioranelli, and H. Griffiths, "Dynamic hand gesture classification based on radar micro-Doppler signatures," in *Proc. CIE Int. Conf. Radar (RADAR)*, Oct. 2016, pp. 1–4.

[16] X. Zhang, Q. Wu, and D. Zhao, "Dynamic hand gesture recognition using FMCW radar sensor for driving assistance," in *Proc. 10th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2018, pp. 1–6.

[17] S. Y. Kim, H. G. Han, J. W. Kim, S. Lee, and T. W. Kim, "A hand gesture recognition sensor using reflected impulses," *IEEE Sensors J.*, vol. 17, no. 10, pp. 2975–2976, May 2017.

[18] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-Doppler signatures with convolutional neural network," *IEEE Access*, vol. 4, pp. 7125–7130, 2016.

[19] Z. Chen, G. Li, F. Fioranelli, and H. Griffiths, "Dynamic hand gesture classification based on multistatic radar micro-Doppler signatures using convolutional neural network," in *Proc. IEEE Radar Conf. (RadarConf)*, Apr. 2019, pp. 1–5.

[20] S. Skaria, A. Al-Hourani, M. Lech, and R. J. Evans, "Hand-gesture recognition using two-antenna Doppler radar with deep convolutional neural networks," *IEEE Sensors J.*, vol. 19, no. 8, pp. 3041–3048, Apr. 2019.

[21] J. S. Suh, S. Ryu, B. Han, J. Choi, J.-H. Kim, and S. Hong, "24 GHz FMCW radar system for real-time hand gesture recognition using LSTM," in *Proc. Asia–Pacific Microw. Conf. (APMC)*, Nov. 2018, pp. 860–862.

[22] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with soli: Exploring fine grained dynamic gesture recognition in the radio-frequency spectrum," in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, Oct. 2016, pp. 851–860.

[23] S. Hazra and A. Santra, "Short-range radar-based gesture recognition system using 3D CNN with triplet loss," *IEEE Access*, vol. 7, pp. 125623–125633, 2019.

[24] F. Khan, S. K. Leem, and S. H. Cho, "In-air continuous writing using uwb impulse radar sensors," *IEEE Access*, vol. 8, pp. 99302–99311, 2020.

[25] S. K. Leem, F. Khan, and S. H. Cho, "Detecting mid-air gestures for digit writing with radio sensors and a CNN," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1066–1081, Apr. 2020.

[26] M. Kurmanji and F. Ghaderi, "A comparison of 2D and 3D convolutional neural networks for hand gesture recognition from RGB-D data," in *Proc. 27th Iranian Conf. Electr. Eng. (ICEE)*, Apr. 2019, pp. 2022–2027.

[27] W. Zhang and J. Wang, "Dynamic hand gesture recognition based on 3D convolutional neural network models," in *Proc. IEEE 16th Int. Conf. Netw., Sens. Control (ICNSC)*, May 2019, pp. 224–229.

[28] I. Fredriksen, *XeThru X4 Radar User Guide*, Novelda, Oslo, Norway, Sep. 2018.

[29] J. D. Taylor, *Introduction to Ultra-Wideband Radar Systems*, 1st ed. Boca Raton, FL, USA: CRC Press, Dec. 1994.

[30] H. A. Hjortland, D. T. Wisland, T. S. Lande, C. Limbodal, and K. Meisal, "Thresholded samplers for UWB impulse radar," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2007, pp. 1210–1213.

[31] J. Li, "Parallel two-class 3D-CNN classifiers for video classification," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Nov. 2017, pp. 7–11.

[32] P. Moradi and M. Jamzad, "Detecting lung cancer lesions in CT images using 3D convolutional neural networks," in *Proc. 4th Int. Conf. Pattern Recognit. Image Anal. (IPRIA)*, Mar. 2019, pp. 114–118.

[33] I. Sahumbaiev, A. Popov, J. Ramirez, J. M. Gorriz, and A. Ortiz, "3D-CNN HadNet classification of MRI for Alzheimer's disease diagnosis," in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf. (NSS/MIC)*, Nov. 2018, pp. 1–4.

[34] H. Williams, L. Cattani, W. Li, M. Tabassian, T. Vercauteren, J. Deprest, and J. D'hooge, "3D convolutional neural network for segmentation of the urethra in volumetric ultrasound of the pelvic floor," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2019, pp. 1473–1476.

[35] T. Guo, J. Dong, H. Li, and Y. Gao, "Simple convolutional neural network on image classification," in *Proc. IEEE 2nd Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2017, pp. 721–724.

[36] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[37] Y. Kim and H. Ling, "Human activity classification based on micro-Doppler signatures using a support vector machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 5, pp. 1328–1337, May 2009.

[38] J.-W. Choi, S.-J. Ryu, and J.-H. Kim, "Short-range radar based real-time hand gesture recognition using LSTM encoder," *IEEE Access*, vol. 7, pp. 33610–33618, 2019.

[39] J. Cifuentes, P. Boulanger, M. T. Pham, F. Prieto, and R. Moreau, "Gesture classification using LSTM recurrent neural networks," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 6864–6867.

[40] Y. Wang, "A new concept using LSTM neural networks for dynamic system identification," in *Proc. Amer. Control Conf. (ACC)*, May 2017, pp. 5324–5329.

SRUTHY SKARIA (Member, IEEE) received the master's degree in electronics engineering, in 2015. She is currently pursuing the Ph.D. degree with the School of Engineering, RMIT University, Melbourne, VIC, Australia. Her current research interests include consumer radars and its applications in the fields of automotive, gesture recognition, UAVs, interference issue in consumer radars, and applications of neural network in radar technology. Her publication Hand-Gesture Recognition Using Two-Antenna Doppler Radar With Deep Convolutional Neural Networks has received the Best Paper Award runner-up from the IEEE Sensors Journal, since 2020.

AKRAM AL-HOURANI (Senior Member, IEEE) received the Ph.D. degree from RMIT University, Melbourne, VIC, Australia, in 2016. He is currently a Senior Lecturer and the Program Manager for the Master of Engineering (Telecommunication and Networks) with the School of Engineering, RMIT University. He has published more than 55 journal articles and conference proceedings, including three book chapters. Since 2020, he has won the IEEE Sensors Council Paper Award for his contribution in hand-gesture recognition using neural networks. He has extensive industry/government engagement as a Chief Investigator in multiple research projects related to the Internet-of-Things (IoT), smart cities, and satellite/wireless communications. As a Lead Chief Investigator, he oversaw the design and deployment of the largest open IoT network in Australia in collaboration with five local governments Northern Melbourne Smart Cities Network, this project has won the 2020 IoT awards, and the official awards program of IoT Alliance Australia. Prior his academic career, from 2006 to 2013, he had extensively worked in the ICT industry sector as an Research and Development Engineer, a Radio Network Planning Engineer, and then as an ICT Program Manager for several projects spanning over different technologies; including mobile networks deployment, satellite networks, and railway ICT systems. His current research interests include UAV communication systems, automotive and mmWave radars, energy efficiency in wireless networks, and the Internet-of-Things over satellite. He is also serving as an Associate Editor for *Frontiers in Space Technologies* and *Frontiers in Communications and Networks*, and as a Guest Editor for the Special Issue Satellite Communication in MDPI Remote Sensing.

ROBIN J. EVANS (Life Fellow, IEEE) received the B.E. degree in electrical engineering from The University of Melbourne, Melbourne, VIC, Australia, in 1969, and the Ph.D. degree from the University of Newcastle, Callaghan, NSW, Australia, in 1975. He is currently a Melbourne University Laureate Professor and a Chief Investigator with the ARC Center of Excellence for Gravitational Wave Detection. His research and industry engagement has ranged across many areas, including theory and applications in control systems, industrial electronics, radar systems, signal processing, and telecommunications. He is also a Fellow of the Australian Academy of Science, the Australian Academy of Technological Sciences and Engineering, and the Institution of Engineers Australia.