# Radar-Based Air-Writing Gesture Recognition Using a Novel Multistream CNN Approach

Shahzad Ahmed, Wancheol Kim, Junbyung Park, and Sung Ho Cho, *Member, IEEE*

***Abstract*—Hand gestures, being a convenient and natural way of communication, is getting huge attention for human–computer interface designs. Among these gestures, detecting mid-air writing is one of the most promising applications. Existing radar-based solutions often perform the mid-air writing recognition by tracking the hand trajectory using multiple monostatic or bistatic radars. This article presents a multistream convolutional neural network (MS-CNN)-based in-air digits recognition method using a frequency-modulated continuous-wave (FMCW) radar. With one FMCW radar comprising of two receiving channels, a novel three-stream CNN network with range-time, Doppler-time, and angle-time spectrograms as inputs is constructed and the features are fused together in the later stage before making a final recognition. Unlike the traditional CNN, MS-CNN with multiple independent input layers enables the creation of a multidimensional deep-learning model for FMCW radars. Twelve human volunteers were invited to writing the digits from zero to nine in the air in both home and lab environments. The three-stream CNN architecture-based air writing for digits has shown a promising accuracy of 95%. A comparison of the proposed MS-CNN system was made with 45 different variants of CNN and preliminary results shows that MS-CNN outperforms the other traditional CNN architectures for air-writing application. The gestures radar data have also been made available to the research community.**

***Index Terms*—Deep learning, frequency-modulated continuous-wave (FMCW) radar, hand gesture recognition, in-air writing, multistream CNN.**

## I. Introduction

RECOGNITION of the voluntary movements of human hands containing a conveyable information is known as hand gesture recognition [1]. Recently, in the field of human–computer interaction (HCI), where the previously available solutions are becoming a bottleneck, hand gesture recognition has widely been studied [2]. One such common application of gesture recognition is the mid-air writing recognition. Researchers have previously utilized various gesture sets for developing HCI systems and among them, the air-writing gesture set is one of the most useful and challenging tasks. The term "air writing" is defined as the movement of hand or fingers in the air with an intent to project the characters or digits [3]. Air writing can bring us a wide range of applications. For instance, in the ongoing COVID-19 pandemic situation, HCI capable of providing noncontact text input will create hindrance to the viral spread. Air writing differs slightly from the traditional writing on a paper. In traditional hand writing, only the pen movement is supposed to convey the information about what digit or character is being written. Contrary to that, for mid-air applications, the trajectory of hand movement expresses the information about what is being written.

Based on the type of sensor being used for data acquisition, gesture recognition systems can largely be classified into two classes [4]: 1) wearable sensor based and 2) wireless sensor based. Wearable sensor requires the user to attach the sensor to their body. For instance, Al-Qaness *et al.* [5] presented activity and gesture recognition using Inertial Measurement Units. For air writing, wearable sensors such as hand held sensors [6] and body worn sensors [7] have shown a promising recognition accuracy. However, these sensors may cause discomfort to the users as users are always required to wear a sensor on the body. Additionally, this type of solution still requires users to physically touch the sensor. Wireless sensors on the other hand, provides more natural way of air-writing implementation. Camera and radio sensors such as radar are the most commonly used wireless sensors. Unlike camera, radar sensors has no associated privacy issue since radar sensor only records the reflections related to the hand movement. This makes radar a suitable candidate for mid-air solutions. Specifically, the complex data cube of a multiinput–multioutput (MIMO) frequency-modulated continuous-wave (FMCW) radar provides a wide range of meaningful information about the target. Contrary to an unmodulated (single frequency) continuous-wave radar, FMCW radar is capable of providing rich information about the target's range, Doppler-velocity and angle simultaneously. Consequently, the FMCW radar has widely been explored previously for several applications, such as vital sign monitoring [8], human gait analysis [9] and specifically, hand gesture recognition [10]. Perhaps, radar has recently shown its footprints for multiple target gesture recognition as well [11]. Nowadays, devices such as Google Pixel 4 smartphone contains in-built radar sensor [12] dedicated solely for gesture recognition-based applications.

Similar to any other field, the use of machine learning for designing the HCI system based on gesture recognition

Manuscript received 27 April 2022; revised 11 June 2022; accepted 30 June 2022. Date of publication 8 July 2022; date of current version 21 November 2022. This work was supported by the National Research Foundation (NRF) of Korea under Grant NRF-2022R1A2C2008783. *(Shahzad Ahmed and Wancheul Kim are co-first authors.) (Corresponding author: Sung Ho Cho.)*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB) on Human Subject Research and Ethics Committee, Hanyang University, Seoul, South Korea, under Application No. HYUIRB-202101-015, and performed in line with the regulations and guidelines suggested by the IRB board.

The authors are with the Department of Electronic Engineering, Hanyang University, Seoul 04763, South Korea (e-mail: dragon@hanyang.ac.kr).

Digital Object Identifier 10.1109/JIOT.2022.3189395

through radar has widely been explored [2], [13]. While designing such HCI systems for radars, the gesture set is often limited, which makes the supervised machine learning as prominent candidate solution. The two main classes of machine-learning-based classification are features based and representation based [14]. Feature-based learning requires a set of handcrafted features; whereas, the representation-based learning approaches operate on the raw data itself [14]. In context of gesture recognition using the FMCW radar, both features extraction-based recognition [15] and the representation-learning-based recognition [16] have previously been considered. Next, the work related to radar sensor-based in-air writing is discussed.

### A. Related Work

As stated earlier, a considerable amount of research work has been presented on noncontact air writing through optical sensors, such as camera [17] and depth camera [18]. Tracking the hand trajectory with camera sensor followed by a detection and recognition algorithm has widely been considered by researchers. For example, Chen *et al.* [3] used a hidden Markov model (HMM) on the tracked trajectory to classify characters. Camera-based finger and hand tracking accompanied by CNN has also been proven as a promising HCI system [19]. Nevertheless, recently, radar-based air-writing systems are gaining huge attention. For instance, Leem *et al.* [20] presented a digit writing system using three Impulse radars where the trilateration-based trajectory estimate was used to draw the hand motion which served as an input to a CNN architecture. Their work was one of the initial attempts for impulse-radars-based air-writing system. However, the proposed system is not appropriate for commercial applications as it always requires three or more radars to recognize the digits. Khan *et al.* [4] presented air writing of digit and few alphabets using an Impulse radar network. Similar to [20], Khan *et al.* also used a CNN architecture driven by the motion sketch drawn by trilateration algorithm. However, the proposed method requires four monostatically configured impulse radars installed at different locations. The main difference between these two aforementioned works [4], [20] was the number of gestures being classified and number of radars being used. With the FMCW radar, Arsalan and Santra [21] used a similar approach of trajectory tracking followed by a temporal CNN. Another work by Arsalan and Santra used 3-D trilateration with deep learning for recognizing five digits and five alphabets. Another work [22] used a tracking algorithm for mid-air writing implementation using FMCW radar. Existing FMCW radar-based air-writing solutions often operates on the hand movement trajectory tracking with the network of radars.

### B. Contribution

As explained earlier, the previous works mostly used hand trajectory tracking followed by a deep learning architecture for classifying the digits or alphabets. A network of radars installed at different locations in data capturing environment is required [4], [20]–[22], making the overall solution a bit expensive and less practical. Furthermore, the previous works relied only on the distance ques. The FMCW radar on the other hand provides target information in different domains, such as range-time, Doppler-time, and angle-time domains. However, single stream CNN with either of these as input may not be suitable for classifying complex tasks such as mid-air digit writing. In this work, we present a new approach for FMCW radar-based air-writing recognition using a multistream CNN model. FMCW radar with two receiving channels only is used in our system to records the movements corresponding to the in-air digit writing activity. As shown in Fig. 1, the system simultaneously extracts the convolutional features from range-time maps (RTMs), Doppler-time maps (DTMs), and angle-time maps (ATMs), and concatenates these features in later stage within the network. With this approach, time variations of distance, the directional velocity, and the angle-of-arrival of the hand collectively contribute to classify different gestures, yielding a compact solution.

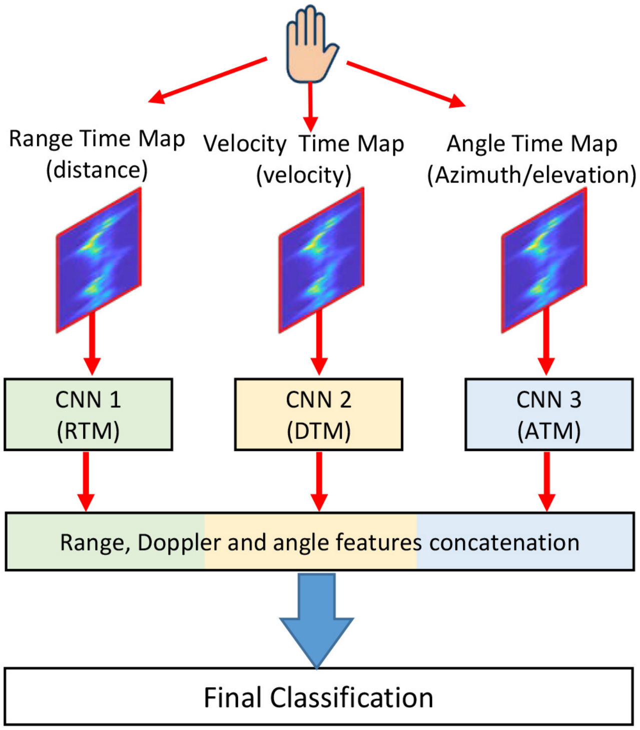


Fig. 1. Multistream CNN model with RTM, DTM, and ATM as input.

In context of deep learning, several variations of Neural Nets have been explored for radar, such as Googlenet [23], temporal CNN [24], and equal-volume (EV) Neural Net [25]. However, the use of the three-stream CNN with the late fusion technique is yet to be explored. There main contributions of our work are as follows.

1) Our work presents a compact solution for air writing the digits using the FMCW radar comprising of two receiving channels only. Previous radar-based air-writing systems do not utilize the range, velocity, and angle simultaneously. To the best of our knowledge, this is the first attempt of air writing using a two-channel FMCW radar. Existing works only use the range ques with localization and tracking algorithms to classify the digits. Consequently, in most of the previous works, a

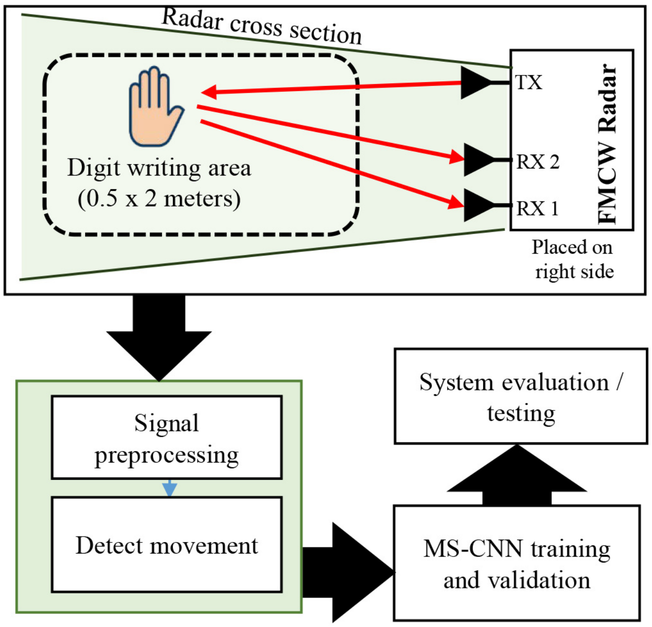


Fig. 2. High-level system block diagram.

network of radar is required, making it less suitable commercially.

2) We present a strategy to implement a multistream CNN for mid-air digit writing. To the best of our knowledge, the three-stream CNN has not yet been explored for Radar-based recognition problems. The proposed architecture can be generalized for any FMCW radar-based classification problem.

3) Unlike the optical sensor where few air-written digit data sets exist such realsense trajectory digits (RTDs) [17], no such radar-based air writing public data set exist. Currently available radar-based hand gesture data sets [2], [24], [25] contains simple gestures such as hand swiping and rotations. A public data set for digits with the radar is disclosed to the research community at the FigShare repository at [26]: https://doi.org/10.6084/m9.figshare.17696435.

## II. Methodology

The overall methodology of digit writing is presented in Fig. 2. A digit is written in the specified area and the corresponding radar returns are preprocessed to de-noise the received signal followed by the gesture duration extraction block. Since, the size of input layer in CNN is usually fixed, time to perform each gesture is fixed to 5 s. Afterward, the RTM, DTM, and ATM images are created, and then fed as input to the deep-learning-based classifier. The radar was installed on the right side to capture more variations in distance. Here, vertical angle (elevation) is calculated using only two receiving channels by exploiting Capon beamforming [27]. Finally, training and test accuracy is computed to access the network performance. Next, we present the details of each block in the further sections.

### A. Radar Signal Preprocessing

The waveform of the signal transmitted by the FMCW radar increases linearly with time, known as a chirp. A single frame comprises usually of one or more such chirps [28]. Upon reflection from hand, the corresponding reflections are received at the receiver antennas. The transmitted signal $x(t)$ having a bandwidth $B$, which can be expressed as [29]

$$x_i(t) = \exp\left(j2\pi\left(f_c t + \frac{B}{T}t^2\right)\right) \tag{1}$$

where the term $f_c$ represents the carrier frequency and $T$ is the duration of pulse. The reflected signal corresponding to a hand located at a distance of $R$, causing a delay of $\tau$ can be expressed as

$$x_r(t) = \exp\left(j2\pi\left(f_c(t-\tau) + \frac{B}{T}(t-\tau)^2\right)\right). \tag{2}$$

The delay $\tau$ depends on the velocity of the target (hand) $v$ and can expressed as

$$\tau = \frac{2(R+v_r t)}{c}. \tag{3}$$

This received signal is mixed with a copy of the transmitted signal and the output of the mixture is a low-frequency signal, termed the IF signal or intermediate frequency signal, such that

$$x_{IF}(t) = \exp\left(j2\pi\left(f_c\tau + \frac{B}{2T}\tau^2\right)\right). \tag{4}$$

### B. RTM, DTM, and ATM Pattern Generation

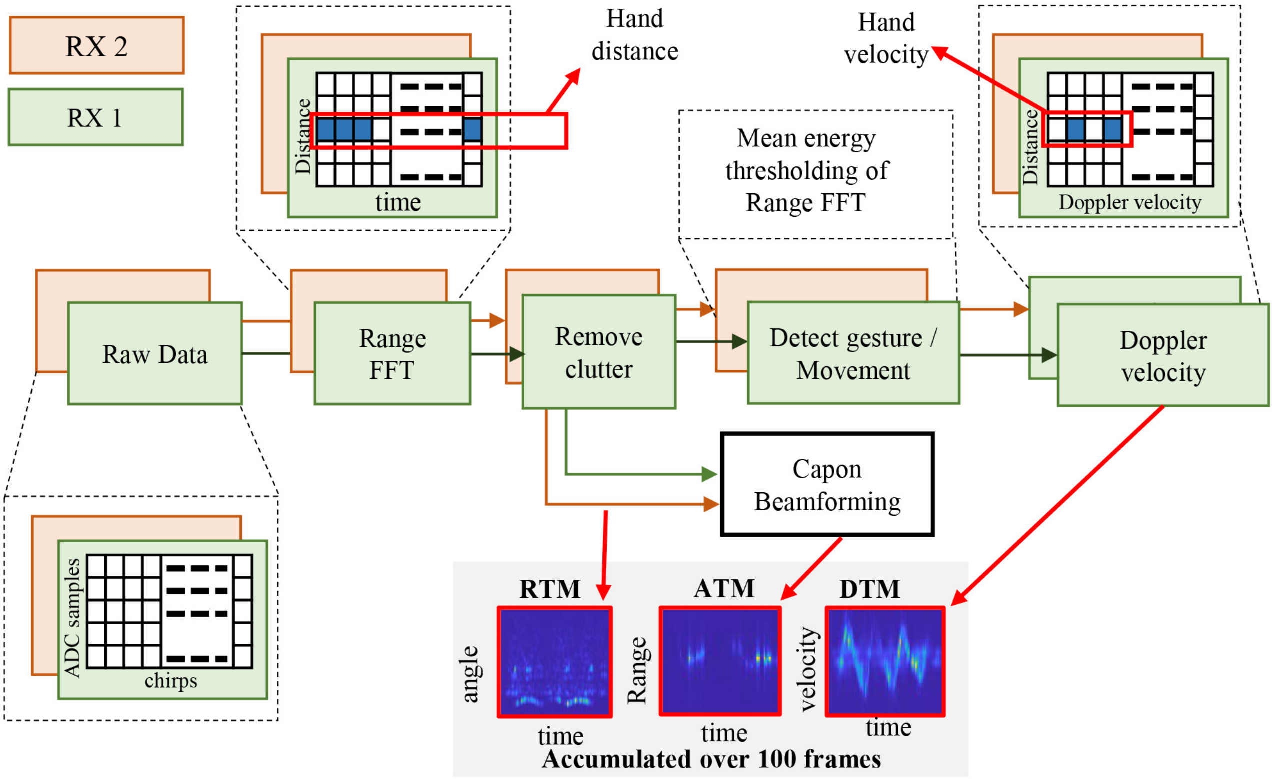
The raw IF signal $x_{\mathrm{IF}}(t)$ contains several chirps. An IF signal containing $N$ chirps can be arranged in the matrix form with each column representing an individual chirp and the row representing all the samples of that chirp to form a 2-D matrix of size $M$ by $N$, expressed as

$$S_{\mathrm{raw}} = \begin{pmatrix} Ch_1[1] & Ch_2[1] & . & Ch_N[1] \\ Ch_1[2] & Ch_2[2] & . & Ch_N[2] \\ . & . & . & \\ Ch_1[M] & Ch_2[M] & . & Ch_N[M] \end{pmatrix} \tag{5}$$

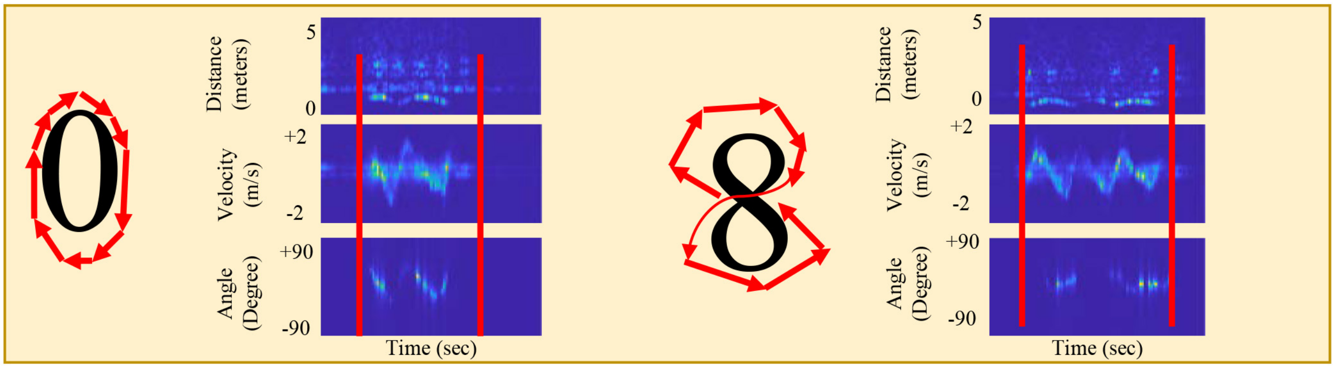
where the term $Ch$ represents the individual chrip and $M$ represents the number of samples for each chirp. The radar raw data matrix expressed in (5) is used to extract the range, velocity, and angle information of hand. A summarized workflow to process the radar frame is shown in Fig. 3(a). The matrix represented in (5) for each receiving channel is shown in the top left corner of Fig. 3(a). For each of the received frame, the data shown in (5) are first denoised by subtracting the frame values with the mean value of the frame. Afterward, each chirp is individually multiplied with the Hamming window function to reduce the effect of side lobes from the data [30]. Fast Fourier transform (FFT) against each chirp in the column of the above-formulated matrix results in the RTM, where the peaks of FFT correspond to the location of the target located within the operational range of the radar

$$S_{\mathrm{FFT1}} = \mathrm{columnFFT}(S_{\mathrm{raw}}). \tag{6}$$

Here, the signal is passed through a loop-back filter [23] for clutter estimation and removal. In order to detect any



(a)



(b)

Fig. 3. (a) FMCW radar signal preprocessing pipeline to generate RTMs, DTMs, and ATMs. (b) Generated maps while writing digit zero and eight in front of the radar.

movement in the clutter-removed radar signal, the mean value of the above matrix is calculated and normalized between 0 and 1 as follows:

$$S_n = \frac{\text{mean}\{\text{mean}(S_{\text{FFT1}})\} - \min\{\text{mean}(S_{\text{FFT1}})\}}{\max\{\text{mean}(S_{\text{FFT1}})\} - \min\{\text{mean}(S_{\text{FFT1}})\}} \quad (7)$$

where $S_n$ represents a 1-D vector ranging between 1 and 0. If the average of 20 consecutive sample values $S_n$ cross the threshold of 0.3, then we say that a movement is detected and Doppler velocity of this observation window is computed. Else, the observation window is time shifted by 20 frames. This detection threshold is selected based on the trial-and-error approach. The Doppler velocity of the target can be computed by taking another FFT against each of row such that

$$S_{\text{FFT2}} = \text{rowFFT}(S_{\text{FFT1}}). \quad (8)$$

Apart from the distance and velocity, another import information related to target (hand) is the angle of arrival. In order to extract the angle, as explained in Fig. 3(a), two receiving channels of radar are utilized. A simplest way to get an angle is to perform third FFT against the peak values of the range–Doppler map across each receiving channels. However, to achieve a higher angle resolution, a beamforming technique must be used. Here, we extracted the angle using the Capon Beamforming technique [31]. In order to calculate the angle, range FFT results corresponding to both the Receiver 1 and the Receiver 2 were exploited using the minimum variance distortionless response (MVDR) algorithm proposed by Capon [31]. The corresponding output of the MVDR beamformer was saved as an angle-time image.

While performing a single-digit writing activity, the RTM, DTM, and ATM images are formed by accumulating the individual images against each frame. Fig. 3(b) represents the RTM, DTM, and ATM generated against digit zero and eight being performed for 5 s. Here, the red lines across the digit represent the trajectory directions. The information corresponding to each gesture is also highlighted as red. A clear distinction in patterns can be observed for these two gestures in terms of velocity and angle maps. Note that the radar was located at the right side of the human volunteer while capturing the digit. Next, as explained earlier, these images serve as input to the deep learning architecture.

### *C. Deep Learning Architecture*

The FMCW radar is capable of providing multidimensional information of target in different domains. The foremost or the basic information is the change in the distance caused by hand movement termed range-time variations or RTM. The RTM can further be processed to extract the (Doppler) velocity and the angle of arrival of hand with respect to the radar. The multistream CNN architecture here is aimed to be capable of extracting features from the available radar data simultaneously from RTM, DTM, and ATM. The basics of a CNN architecture are defined here.

*1) CNN Architecture:* A traditional CNN architecture comprises mainly of three layers.

1) *Input Layer:* In this layer, the radar data in the form of images are fed into the convolutional network. The data can either be 2-D in the form of grayscale images, or the 3-D red, green, and blue (RGB) images. For the FMCW radar as stated earlier, there are a wide range of available data that can serve as the input layer, such as range-time, Doppler frequency-time, angle-time, range-angle, and so on. The traditional CNN can consider either of these data as input.
2) *Hidden Layer:* The hidden layer can further consist of few layers, such as the convolutional layer, batch-normalization layer, max-pooling layer, and rectified linear unit ReLU layer. Among them, the convolutional layer is the core of a CNN. Here, the input image or the output of the previous hidden layer is convolved with a predefined filter, commonly termed kernel. The multidimensional convolution is performed by sliding the kernel through the entire image. The convolutional operation can be defined as

$$\text{Conv}_{\text{out}}[m,n] = \sum_j \sum_k h\big[j,k\big] i_i n\big[m-j, n-k\big] \quad (9)$$

where $h$ represents the kernel filter, $i_{\text{in}}$ represents the input to convolutional layer, and $\text{conv}_{\text{out}}$ represents the output. Following the convolutional operations, batch-normalization is performed which stabilizes the training process and reduces the required epoch sizes. Afterward, a ReLU and max-pooling functions are usually applied. With ReLU acting as activation function, the convolution features are subjected to below operation:

$$\text{ReLU}_{\text{out}} = \max(0, \text{conv}_{\text{out}}(i)) \quad (10)$$

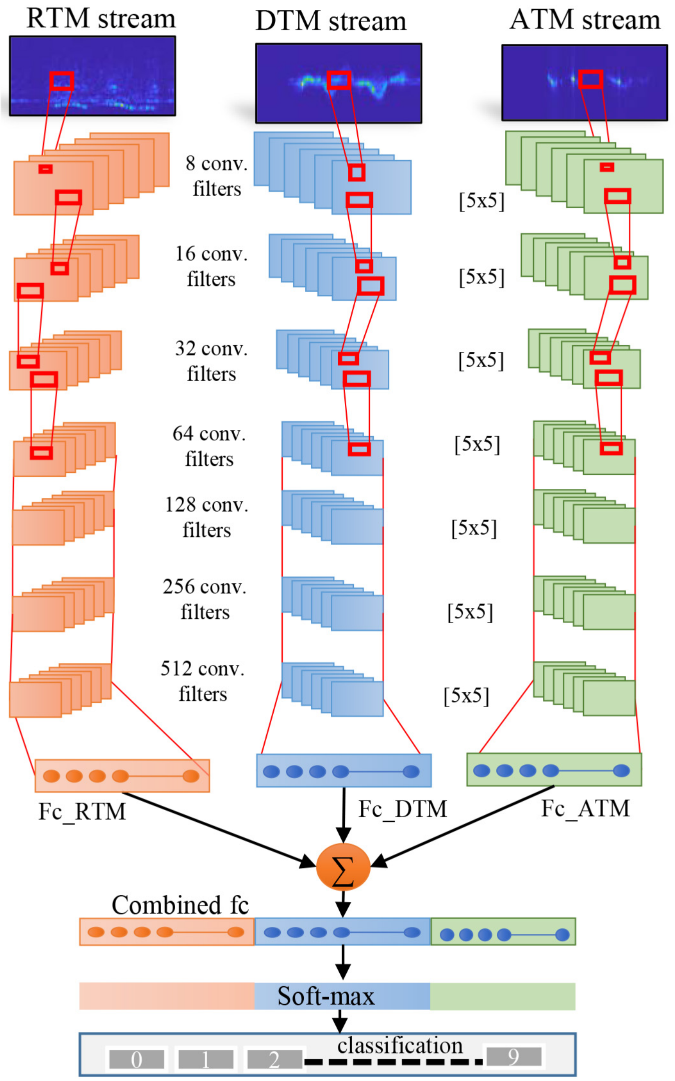


Fig. 4. Architecture of the proposed Multistream CNN model with RTM, DTM, and ATM as input stream.

where $i$ represents the index of the output of the convolutional layer. Finally, the max-pooling layer reduces the size of the convolutional output and performs the dimensionality reduction.

3) *Classification Layer:* After passing the input to several hidden layers, finally, classification is performed. Here, soft-max [32] normalization is often performed which plays a vital role in a multiclass classification problem [32].

*2) MS-CNN:* The opted multistream convolutional neural network (MS-CNN) architecture is shown in Fig. 4. In this architecture, RTM, DTM, and ATM are simultaneous fed to the network as input. Consequently, the relevant features can be extracted independently, creating the nonlinear combination of the extracted features. As shown in Fig. 4, all the three streams comprise of seven hidden layers. Here, each hidden layers contains a convolutional layer, a batch-normalization layer, ReLu layer, and a max-pooling layer. A fixed kernel size of 5×5 is used in each layer for convolution. The number of filters in the first hidden layer is 8 which keeps on increasing in

TABLE I
TI-IWR6843 FMCW RADAR SETTING FOR DATA ACQUISITION

| Radar Parameter | Value / description |
|---|---|
| Starting frequency | 60GHz |
| Total bandwidth Span | 4GHz |
| Number of chirps | 128 |
| Frame rate | 20 frames per second |
| Maximum Range | 11 meters |
| ADC samples per chirp | 100 samples |
| Chirp slope | 60 MHz per micro second |

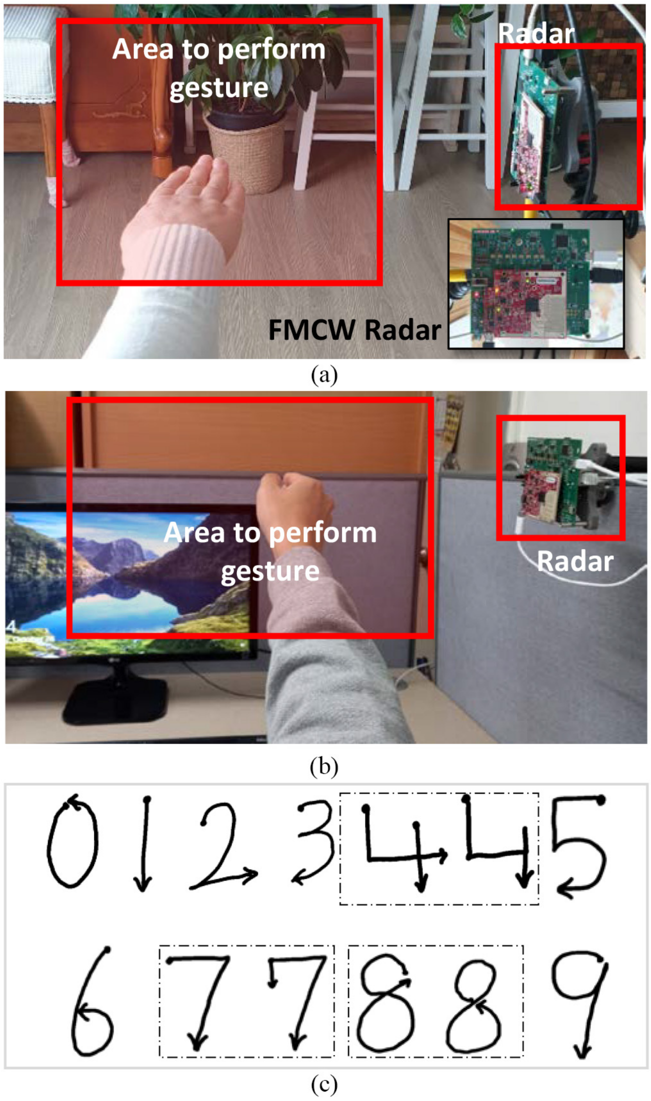
each successive hidden layer and at the final layer, 512 filters are used. Dimensionality reduction is achieved by deploying a max-pooling layer of size 2 × 2 at each hidden layer which reduces the dimension of the next hidden layer by a factor of 2. After passing the radar data through seven hidden layers, a fully connected (fc) layer is formed at each individual stream. The layers named as $fc_{\text{RTM}}$, $fc_{\text{DTM}}$, and $fc_{\text{ATM}}$ contains the features extracted from the time variations of distance, velocity and angle. Finally, the $fc_{\text{combined}}$ layer concatenates the independently extracted convolutional features extracted from RTM, DTM, and ATM patterns. Afterward, the final classification is performed using a softmax decision at the end.

The size and the parameters of the network shown in Fig. 4 are chosen solely on trial-and-error-based experimental evaluation. After testing several structural variations, the best suited architecture is being used for classification purpose. Furthermore, Section III contains the detailed analysis regarding the accuracy of different MS-CNN models.

## III. Experimental Setup

### A. Radar Placement

The air-writing data acquisition setup is shown in Fig. 5. Data were captured at two different locations to make the system robust against different environments. The environment shown in Fig. 5(a) is a home environment whereas the environment shown in Fig. 5(b) is located in the IT-BT building, Hanyang University, Seoul, South Korea. The red highlighted areas represent the space where the participants performed the digit-writing gestures. The involved participants selected their own way of performing the digit gestures and all the observed trajectory paths are shown in Fig. 5(c). For data collection, we used TI-FMCW IWR6843 ISK Radar sensor designed by Texas Instruments (TI) Inc., USA, shown in Fig. 5. The radar offers 4-GHz bandwidth ranging from 60 to 64 GHz which has a short wavelength and often referred to as millimeter-wave (mm-wave) technology. Short wavelength offers high target resolution; however, the indoor path loss is very high [33]. Our system comprises of s single FMCW radar device and multiradar environment may require an additional interference cancelation approach based on multiplexing [34]. Rest of the technical specifications of radar and the opted radar settings for this experiment are described in Table I. Fig. 5 also shows that the radar sensor is attached to an FPGA kit (DCA-1000, TI) which is further connected to the host computer via a local-area connection to capture and process the data for recognition of the drawn digits.



(a)

(b)

(c)

Fig. 5. Data acquisition in (a) home and (b) lab environment. (c) Movement trajectory for all gestures.

Before starting the data collection process, we considered three different radar positions which are left, right, and front sides of the human participant. Radar placement and the corresponding time-Doppler and range-time patterns for digit five are shown in Fig. 6. As expressed in Fig. 6, radar installed at the either (left or right) side of the human volunteer resulted in a more prominent pattern in comparison to the radar installed at the front side of the human volunteer. Higher amount of interclass pattern variations can be seen by installing the radar at either side of the human subject. This is due to the fact that while performing gesture with the radar on the side of a human volunteer, the amount of change in distance between the radar and the hand is significantly higher in comparison to the radar installed in front of the human volunteer. Consequently, the right-side position was selected to create the data set. However, the radar placement can be considered as a design parameter

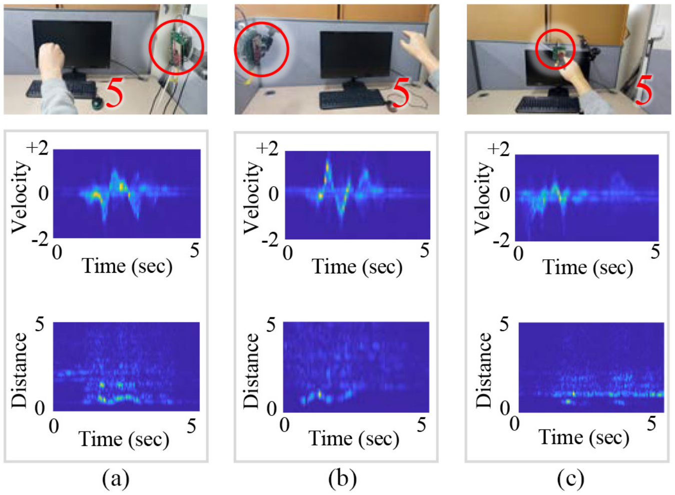


Fig. 6. DTM and RTM patterns for digit “FIVE” when the radar is placed at (a) right, (b) left, and (c) front of a human volunteer.

TABLE II
PARTICIPANT DETAILS

| Participant | Height | Weight | BMI | Age | LH / RH |
|---|---|---|---|---|---|
| 01 - Male | 171 | 69 | 23.5 | 22 | RH |
| 02 - Female | 163 | 53 | 19.9 | 26 | RH |
| 03 - Male | 172 | 67 | 22.6 | 24 | RH |
| 04 - Male | 180 | 85 | 26.2 | 29 | RH |
| 05 - Female | 164 | 59 | 21.9 | 22 | RH |
| 06 - Male | 170 | 74 | 25.6 | 52 | RH |
| 07 - Male | 189 | 94 | 26.31 | 25 | RH |
| 08 - Female | 160 | 57 | 22.26 | 49 | RH |
| 09 - Male | 170 | 67 | 23.1 | 21 | LH |
| 10 - Female | 164 | 73 | 27.1 | 20 | LH |
| 11 - Male | 165 | 68 | 24.9 | 31 | LH |
| 12 - Male | 167 | 62 | 22.2 | 27 | LH |

and can be adjusted accordingly at the beginning of the data collection process. Later, to validate the fact that the proposed algorithm is equally effective when the radar is on the left side, the CNN design shown in Fig. 4 was additionally trained for the radar on the left side of the volunteers. The corresponding results are explained in Section V. Fig. 6(a) and (b) also suggests that for the same gesture, the patterns for the radar on left and right sides appeared to be the flipped version of each other while writing digit five. However, this cannot be generalized since few gestures such as digit one has the same pattern for both positions.

### B. Data Set

To introduce sufficient variations and reduce the biases in the collected data set, 12 participants were involved in the data acquisition exercise having an average age of 29 years with BMI 23.84. Out of the involved volunteers, eight participants were male, whereas, four participants were female. Human data were captured under the guidance of the local ethics committee. Both left and right-handed participants were involved in the data collection process. Rest of the details of each participant are listed in Table II. All the participants were asked to sign an informed consent form prior to the start of the experiment and no identifiable personal information is included in the manuscript. Participants performed the air-writing activity in the specified area and no restrictions, other than being in the designated area, were imposed while collecting the data. Totally, 100 air-writing samples were collected from each participant. For convenience, the file structure of the associated data is shown in Fig. 7. For each participant, there exist ten MATLAB format (MAT) files corresponding to each gesture (0–9). Within a single MAT file, each row contains one FMCW radar frame as expressed in Fig. 7. Here, 100 frames constitute one gesture sample. For instance, to visualize the range-FFT of one gesture, 100 consecutive frames should be loaded in a variable and an FFT across each row will provide a range-FFT map of that specific gesture. Each frame here can be processed using the data processing flow shown in Fig. 3(a) to extract the range, Doppler, and angle features. An additional data visualization code is also placed in the repository.

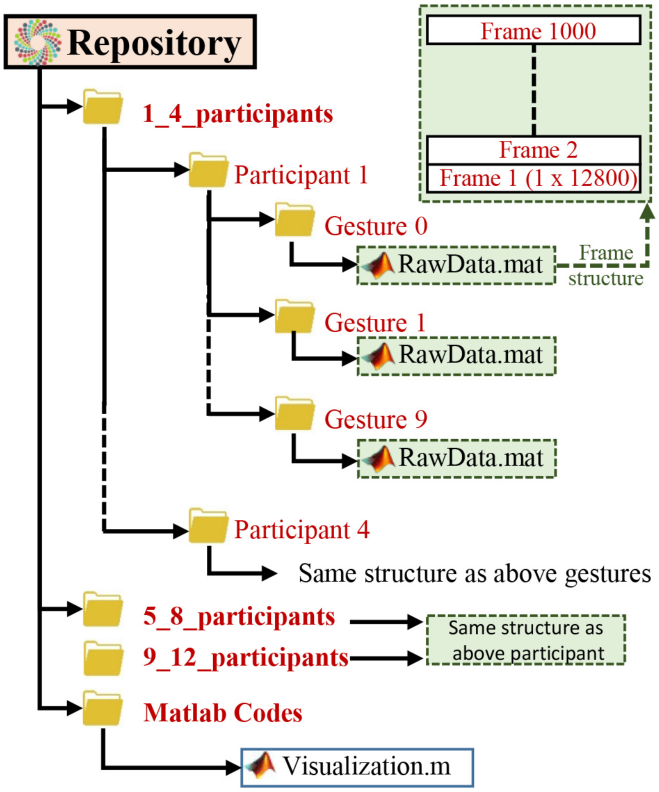


Fig. 7. Structures of files available in the hand gesture data repository.

## IV. EXPERIMENTAL RESULTS

This section shows the experimental results in light of the above-formulated experimental setup.

### A. Movement Detection and Digits Pattern Visualization

Fig. 8 shows the normalized movement index calculated with the Range-FFT signal as explained in (7). As shown in Fig. 8, while writing the digit zero, duration where the digit writing was performed has significantly high magnitude in comparison to the idle time. The RTM, DTM, and ATM

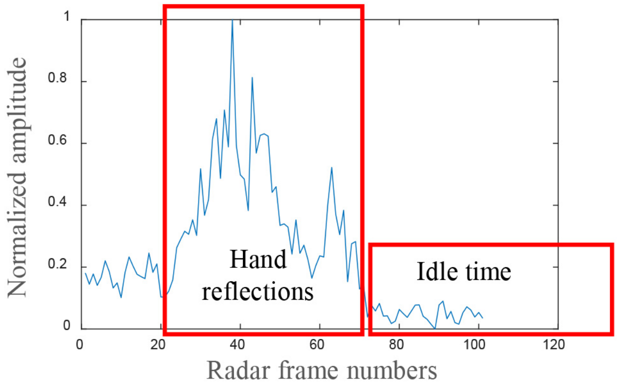


Fig. 8. Normalized range-FFT pattern showing the hand movement duration and the time when no movement was performed.

of all the ten digits starting from zero to nine when the radar was placed on the (right) side of human volunteers are shown in Fig. 9. Note that each gesture was performed for 5 s with frequency of 20 frames/ss. This yields a slow-time (horizontal) axis consisting of 100 observation samples. The vertical axis on the RTM, DTM, and ATM spectrograms, respectively, represents the distance, velocity, and angle values. As seen in Fig. 9, the velocity values are symmetric around the middle point of the velocity axis where the values above the center line represents the target going away from the radar and values below represents the target moving toward the radar. While observing the DTM pattern against digit zero, the velocity pattern can be seen altering the direction couple of times. Fig. 9 suggests that the DTM plots for each digit contains more variation in the patterns of digit in comparison to the RTM and ATM plots.

### B. Intragesture Variation Analysis

Intraclass variations are defined as the variations occurring among different samples of the same class (label) [35]. While the preceding section demonstrated the intraclass variations, this section demonstrates the variations among different samples of the same gesture in the context of hand speed. As stated earlier, in the interest of robustness, no additional restriction regarding the hand speed was imposed on the participants. Fig. 10 shows the speed variations with time while writing "Eight" in the air. The speed in Fig. 10(a) is the slowest (among these four gestures) whereas Fig. 10(d) corresponds to the fastest speed. In the training data, it was observed that participant 8 (Female, 49) performed gestures slowly, whereas participant 5 (Female, 22) performed gestures quickly in comparison to the other participants. Every sample in the data set has its own associated hand speed.

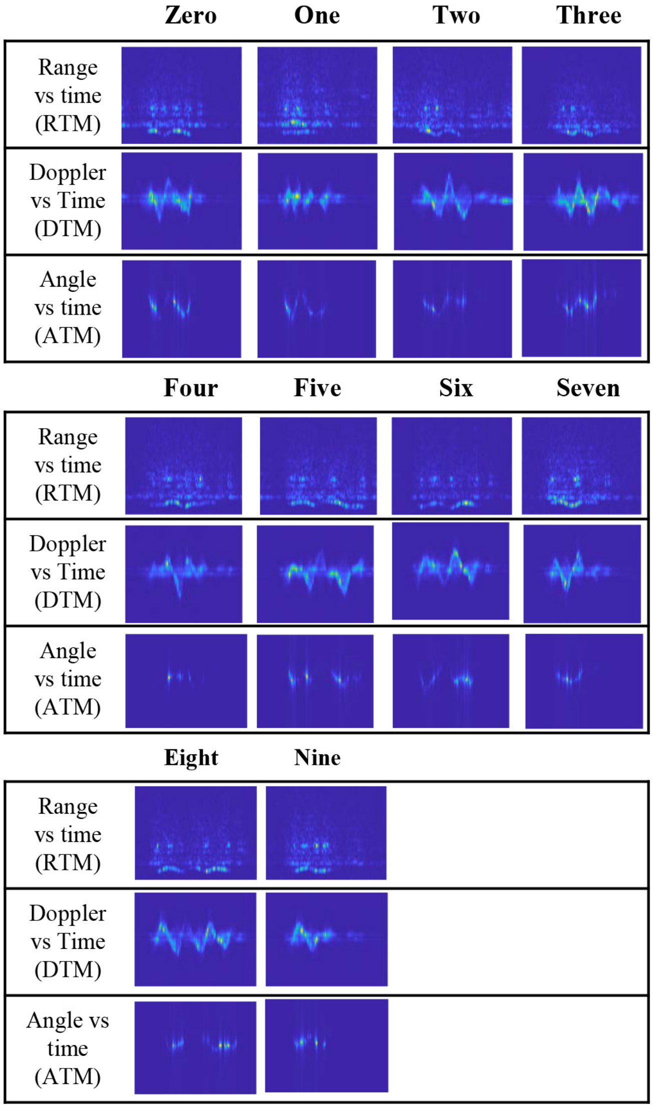


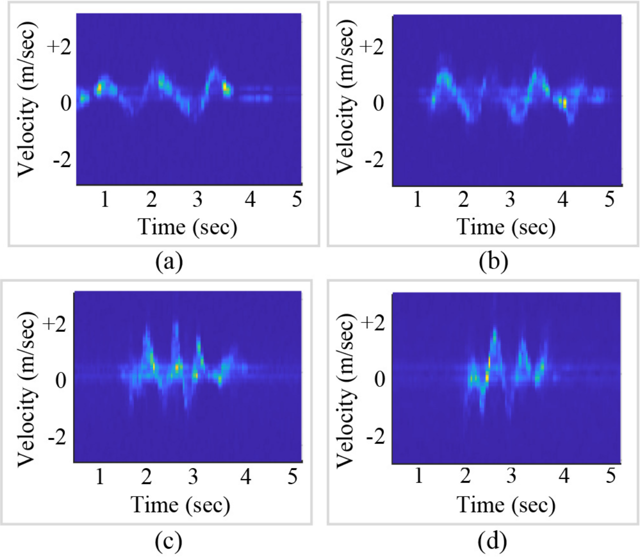
Fig. 9. Time-varying range (RTM), Doppler-velocity (DTM), and angle (ATM) patterns for digits ranging from zero to nine.

TABLE III
ACCURACY OF DIFFERENT MS-CNN VARIANTS AGAINST DIFFERENT NUMBER OF HIDDEN LAYERS AND CONVOLUTIONAL FILTERS

| Filter Size | 3 Layers | 4 layers | 5 layers | 6 layers | 7 layers |
|---|---|---|---|---|---|
| 3x3 | 61.11 | 80.00 | 85.00 | 84.45 | 92.71 |
| 5x5 | 62.78 | 81.67 | 86.67 | 90.56 | 94.20 |
| 7x7 | 70.56 | 87.22 | 87.78 | 91.67 | 90.00 |

### C. Classification Accuracy

CNN-based classifiers have shown a huge success in different applications. The most commonly used CNN architecture contains a combination of convolutional, max-pooling, normalization, and fc layers. Despite being in use extensively, selecting the network size and hyperparameters is still a tedious task [23], [36], [37]. We selected the network using the trial-and-error method to search for the best suited MS-CNN architecture for in-air digit writing recognition. Several combinations of deep-learning architectures with different hyperparameters were analyzed while evaluating the different available networks. The network structure was tuned and an optimized set of hyperparameters was selected. The classification task was repeated several times by varying number of hidden layers and the Kernel (convolution) filter size. Table III shows the detailed results of the performed structural variations. The first column represents the filter size, whereas the first row represents the number of hidden layers. Throughout



(a) (b) (c) (d)

Fig. 10. (a)–(d) Variations in hand movement while performing gesture eight with (a) being slowest, and (d) being highest among four samples under consideration.

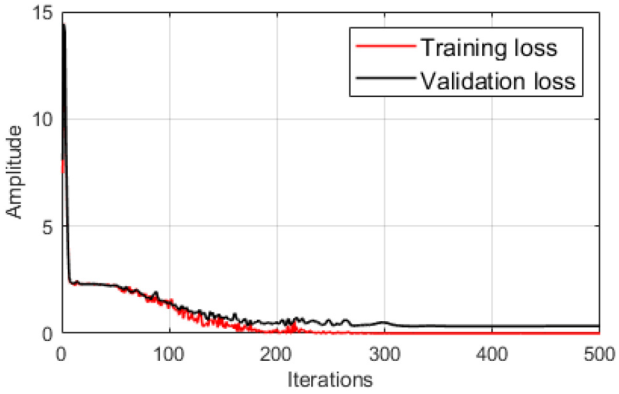
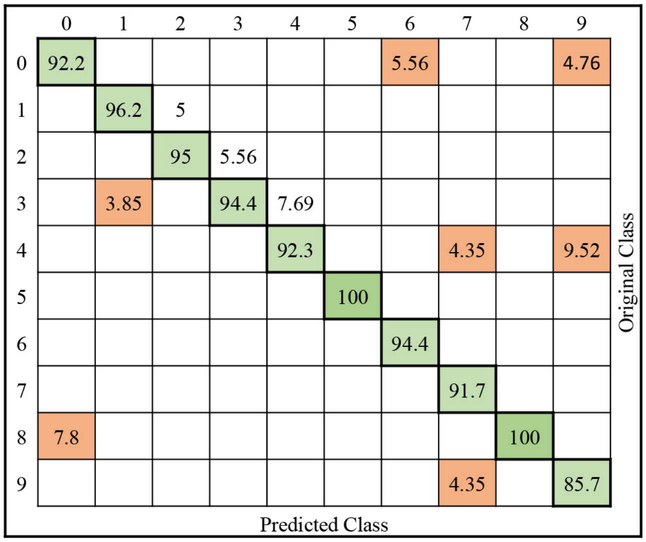


Fig. 11. Training and validation progress of the opted model for 500 iterations.

these experiments, the obtained classification accuracy ranges between 61.11% and 94.20%. The network shown in Fig. 4, comprising of seven layers and "5 × 5" convolutional filter showed maximum accuracy. The training loss and the validation loss for this network are shown in Fig. 11. Note that the data were split into 70%, 15%, and 15% for training, validation, and test purposes. The confusion matrix of the test data for the opted deep-learning model is shown in Fig. 12. The vertical axis represents the known true class of the gesture, whereas the horizontal axis represents the predicted class of the gesture. The values in the diagonal show the classification accuracy of the digit gesture. In addition to the overall accuracy, precision, recall and, $F1$score are also crucial factors to evaluate the effectiveness of any trained model. Precision is calculated based on the true-positive (TP) and false-positive (FP) predictions and defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{11}$$



| Original Class | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 92.2 | | | | | | 5.56 | | | 4.76 |
| 1 | | 96.2 | 5 | | | | | | | |
| 2 | | | 95 | 5.56 | | | | | | |
| 3 | | 3.85 | | 94.4 | 7.69 | | | | | |
| 4 | | | | | 92.3 | | | 4.35 | | 9.52 |
| 5 | | | | | | 100 | | | | |
| 6 | | | | | | | 94.4 | | | |
| 7 | | | | | | | | 91.7 | | |
| 8 | 7.8 | | | | | | | | 100 | |
| 9 | | | | | | | | 4.35 | | 85.7 |

Predicted Class

Fig. 12. Confusion matrix of test data which were not used for the training and validation.

TABLE IV
FORTY FIVE DIFFERENT EXPERIMENTAL SCENARIOS FOR COMPARATIVE ANALYSIS

| Varying quantity | Number of variations (Details) |
|---|---|
| Input to CNN | 3 (RTM, DTM and ATM images) |
| Hidden Layers | 5 (3,4,5,6, and 7 layers) |
| 3 (3x3, 5x5, and 7x7) | |
| Total Variations | 45 (3x5x3) |

On the other hand, recall is being calculated based on TP and false-negative (FN) prediction and defined as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{12}$$

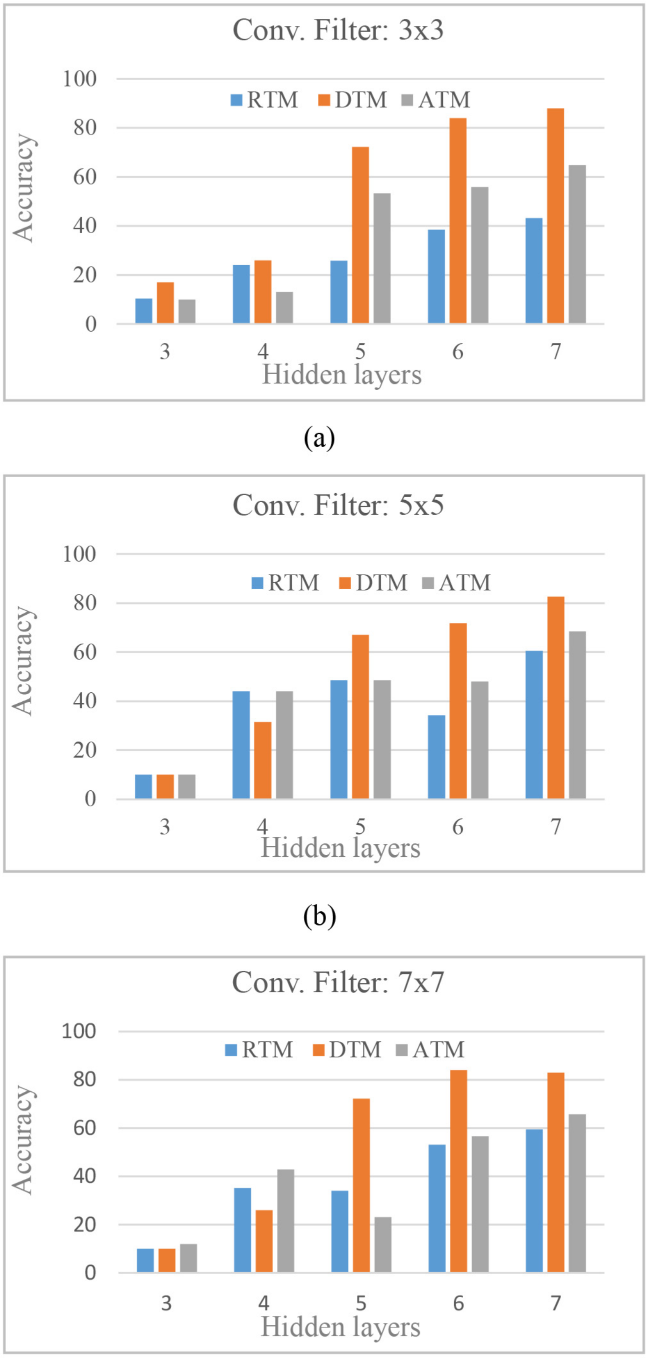
As expressed in the confusion matrix shown in Fig. 12, the precision and recall for the proposed network are 0.9416 and 0.9435. The $F1$score that computes the balance between the precision and recall can be expressed using (11) and (12)

$$F1\text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{13}$$

For the presented model, the $F1$score is 0.9425. High values of precision, recall, and $F1$score were observed which further validates the effectiveness of the model.

### D. Comparison With State of the Art

For comparative analysis, we compare the proposed MS-CNN-based digit-recognition with the traditional single-input CNN. We tested the 45 different CNN architectures (having a single-input layer) with the proposed MS-CNN architecture. The details of the designed experiments for comparative analysis are summarized in Table IV. With three different filter sizes of 3 × 3, 5 × 5, and 7 × 7, CNN with three, four, five, six, and seven hidden layers were created. All these networks were separately fed with RTM, DTM, and ATM as input for



(a)

(b)

(c)

Fig. 13. Digit classification with different traditional CNN architectures having only RTM, DTM, and ATM images as input. Accuracy of CNN having 3–7 hidden layers with (a) 3 × 3 (b) 5 × 5, and (c) 7 × 7 convolution filters.

TABLE V
PERFORMANCE EVALUATION IN DIFFERENT ENVIRONMENTAL CONDITIONS

| Variations | Accuracy (%) |
|---|---|
| Radar on left side | 93.1 |
| Radar on right side | 94.20 |
| Right Handed Participants only | 95.00 |

performance evaluation. Accuracies of all 45 test case scenarios were computed and compared with the proposed MS-CNN architecture. Results of these comparative analysis are summarized in Fig. 13(a)–(c) where the vertical axis represents the accuracy in percentage and the horizontal axis shows the number of hidden layers for a specific filter size. Accuracy for DTM, RTM, and ATM is shown in three different colors. Since the patterns shown earlier in Fig. 8 suggest that while writing digits, the variations in the range as well as the angle with respect to time are not significant enough to distinguish the involved complex movements. As a result, the maximum accuracy was perhaps less than 70% for all the CNN architectures with RTM and ATM input. On the other hand, as shown in Fig. 13(a)–(c), the highest accuracy achieved by training the CNN architecture with DTM as input was 87%. For traditional CNN, DTM as input shows better performance in comparison to RTM and ATM images. In short, with all these 45 experiments, the maximum accuracy of 87% can be achieved. MS-CNN, on the other hand, can achieve classification accuracy of 94.2%. It is evident that the MS-CNN-based air-writing recognition system outperforms the traditional CNN with a single input.

Since the existing radar-based in-air writing techniques work on trilateration methods [4], [20]–[22], hence cannot be implemented with our design methodology since those techniques require multiple radars for tracking the hand gestures. Consequently, a direct comparison of our technique with these existing systems is not possible.

### E. Performance Evaluation in Different Conditions

As explained in Section III, we additionally evaluated the proposed algorithm with the radar placed on the left side, as well and the corresponding accuracy is reported in Table V. Significantly, high accuracy was observed for both the cases of the radar being installed at the left or right side of the human volunteer. This additionally validates the effectiveness of the MS-CNN-based classifier. In addition, we also evaluated the network with right-handed participants (participants 1–8) only. The accuracy was slightly improved when all participants were right handed.

## V. DISCUSSION AND CONCLUSION

In this study, we have introduced a new implementation of in-air digit recognition using the FMCW radar sensor. A multistream CNN model capable of extracting information from the range-time, Doppler-time, and angle-time patterns was proposed. The MS-CNN model combines different features from multiple input streams simultaneously and concatenates the features at the later stage that results in an overall better performance in comparison to the tradition CNN approaches. To introduce diversity and reduce the biasness, data were captured from 12 different participants at different physical environments. Preliminary experimental results have shown that high classification accuracy of 94.20% for recognizing all the ten base digits. The traditional CNN operating on the range-time and Doppler patterns only showed less accuracy in comparison to the MS-CNN. Considering the high accuracy of

MS-CNN, the methodology can also be generalized for gesture recognition problems other than in-air digit writing.

One of the biggest challenges in acquiring features of each gesture is to remove the unwanted noise while keeping the hand reflections intact. Due to the nonrigid structure of the hand, fingers, palm, and other parts of a human hand introduce several additional vibrations known as the micro-Doppler effect. While designing the signal processing framework for filtering the noise, clutter, and the ghost targets, the filters must be designed carefully to retain an adequate amount of micro-Doppler information of the hand. In other words, the filter should remove the unwanted noise only while permitting the hand and the associated micro-Doppler information. Another key issue is the computation cost of a deep learning model. Despite considering different physical environments, the classification accuracy was sufficiently high, however, at the cost of computation burden as the latency of the CNN ranged between 400 and 500 ms.

The in-air writing system proposed in this study is capable of classifying a single digit at a time. This implies that the proposed algorithm treats each performed digit gesture in a discrete fashion. The recognition of continuous digit writing is yet to be explored. In addition, to generate a distinctive pattern for each gesture, users are required to rotate full hand while writing the digit. Finger-tracking-based digit recognition was not investigated in this work.

In the future, we aim to make a contactless in-air writing system for continuous digit recognition and alphabet recognition. In addition to that, we are also aiming to implement a real-time version of the proposed air-writing recognition system and extend MS-CNN-based classification approach to other similar classification problems.

## Acknowledgment

The authors are thankful to the human volunteers who participated in this study.

## References

[1] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: A survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 1–54, 2015.

[2] S. Ahmed, K. D. Kallu, S. Ahmed, and S. H. Cho, "Hand gestures recognition using radar sensors for human–computer–interaction: A review," *Remote Sens.*, vol. 13, no. 3, p. 527, 2021.

[3] M. Chen, G. AlRegib, and B.-H. Juang, "Air-writing recognition—Part I: Modeling and recognition of characters, words, and connecting motions," *IEEE Trans. Human–Mach. Syst.*, vol. 46, no. 3, pp. 403–413, Jun. 2016.

[4] 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.

[5] M. A. A. Al-Qaness, A. Dahou, M. A. Elaziz, and A. M. Helmi, "Multi-ResAtt: Multilevel residual network with attention for human activity recognition using wearable sensors," *IEEE Trans. Ind. Informat.*, early access, Apr. 8, 2022, doi: 10.1109/TII.2022.3165875.

[6] P. Roy, S. Ghosh, and U. Pal, "A CNN based framework for unistroke numeral recognition in air-writing," in *Proc. IEEE 16th Int. Conf. Front. Handwriting Recognit. (ICFHR)*, 2018, pp. 404–409.

[7] P. Kumar, J. Verma, and S. Prasad, "Hand data glove: A wearable real-time device for human–computer interaction," *Int. J. Adv. Sci. Technol.*, vol. 43, pp. 15–25. Jan. 2012.

[8] S. Yoo *et al.*, "Radar recorded child vital sign public dataset and deep learning-based age group classification framework for vehicular application," *Sensors*, vol. 21, no. 7, p. 2412, 2021.

[9] P. Addabbo, M. L. Bernardi, F. Biondi, M. Cimitile, C. Clemente, and D. Orlando, "Gait recognition using FMCW radar and temporal convolutional deep neural networks," in *Proc. IEEE 7th Int. Workshop Metrol. AeroSp. (MetroAeroSpace)*, 2020, pp. 171–175.

[10] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Short-range FMCW monopulse radar for hand-gesture sensing," in *Proc. IEEE Radar Conf. (RadarCon)*, 2015, pp. 1491–1496.

[11] Z. Yu, D. Zhang, Z. Wang, Q. Han, B. Guo, and Q. Wang, "SoDar: Multitarget gesture recognition based on SIMO doppler radar," *IEEE Trans. Human–Mach. Syst.*, vol. 52, no. 2, pp. 276–289, Apr. 2022.

[12] J. Lien *et al.*, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Trans. Graphics*, vol. 35, no. 4, pp. 1–19, 2016.

[13] Y. Zhang, Z. Yang, G. Zhang, C. Wu, and L. Zhang, "XGest: Enabling cross-label gesture recognition with RF signals," *ACM Trans. Sensor Netw.*, vol. 17, no. 4, pp. 1–23, 2021.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] S.-J. Ryu, J.-S. Suh, S.-H. Baek, S. Hong, and J.-H. Kim, "Feature-based hand gesture recognition using an FMCW radar and its temporal feature analysis," *IEEE Sensors J.*, vol. 18, no. 18, pp. 7593–7602, Sep. 2018.

[16] Z. Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor," *IEEE Sensors J.*, vol. 18, no. 8, pp. 3278–3289, Sep. 2018.

[17] C.-H. Hsieh, Y.-S. Lo, J.-Y. Chen, and S.-K. Tang, "Air-writing recognition based on deep convolutional neural networks," *IEEE Access*, vol. 9, pp. 142827–142836, 2021.

[18] M. S. Alam, K.-C. Kwon, and N. Kim, "Implementation of a character recognition system based on finger-joint tracking using a depth camera," *IEEE Trans. Human–Mach. Syst.*, vol. 51, no. 3, pp. 229–241, Jun. 2021.

[19] S. Mukherjee, S. A. Ahmed, D. P. Dogra, S. Kar, and P. P. Roy, "Fingertip detection and tracking for recognition of air-writing in videos," *Exp. Syst. Appl.*, vol. 136, pp. 217–229, Dec. 2019.

[20] 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.

[21] M. Arsalan and A. Santra, "Character recognition in air-writing based on network of radars for human–machine interface," *IEEE Sensors J.*, vol. 19, no. 19, pp. 8855–8864, Oct. 2019.

[22] M. Arsalan, A. Santra, and V. Issakov, "Radar trajectory-based air-writing recognition using temporal convolutional network," in *Proc. 19th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, 2020, pp. 1454–1459.

[23] S. Ahmed and S. H. Cho, "Hand gesture recognition using an IR-UWB radar with an inception module-based classifier," *Sensors*, vol. 20, no. 2, p. 564, 2020.

[24] M. Scherer, M. Magno, J. Erb, P. Mayer, M. Eggimann, and L. Benini, "TinyRadarNN: Combining spatial and temporal convolutional neural networks for embedded gesture recognition with short range radars," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10336–10346, Jul. 2021.

[25] H. Liu *et al.*, "M-gesture: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar," *IEEE Internet Things J.*, vol. 9, no. 5, pp. 3397–3415, Mar. 2022.

[26] S. Ahmed and S. H. Cho. "Radar based air-writing gesture recognition using multi-stream convolutional neural network. figshare. dataset." [Online]. Available: https://doi.org/10.6084/m9.figshare.17696435 (Accessed: Apr. 25, 2022).

[27] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[28] Q. Shi, K. Bunsen, N. Markulic, and J. Craninckx, "A self-calibrated 16-Ghz subsampling-PLL-based fast-chirp FMCW modulator with 1.5-GHz bandwidth," *IEEE J. Solid-State Circuits*, vol. 54, no. 12, pp. 3503–3512, Dec. 2019.

[29] V. Winkler, "Range doppler detection for automotive FMCW radars," in *Proc. IEEE Eur. Radar Conf.*, 2007, pp. 166–169.

[30] F. D. W. Enggar, A. M. Muthiah, O. D. Winarko, O. N. Samijayani, and S. Rahmatia, "Performance comparison of various windowing on FMCW radar signal processing," in *Proc. Int. Symp. Electron. Smart Devices (ISESD)*, 2016, pp. 326–330.

[31] M.-S. Lee and Y.-H. Kim, "Design and performance of a 24-Ghz switch-antenna array FMCW radar system for automotive applications," *IEEE Trans. Veh. Technol.*, vol. 59, no. 5, pp. 2290–2297, Jun. 2010.

[32] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. ICML*, vol. 2, 2016, p. 7.

[33] M. D. N. Anjum, H. Wang, and H. Fang, “Prospects of 60 GHz mmWave WBAN: A PHY-MAC joint approach,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6153–6164, Jun. 2020.

[34] C. Aydogdu, N. Garcia, L. Hammarstrand, and H. Wymeersch, “Radar communications for combating mutual interference of FMCW radars,” in *Proc. IEEE Radar Conf. (RadarConf)*, 2019, pp. 1–6.

[35] M. Taskiran and N. Kahraman, “Comparison of CNN tolerances to intra class variety in food recognition,” in *Proc. IEEE Int. Symp. Innov. Intell. Syst. Appl. (INISTA)*, 2019, pp. 1–5.

[36] M. Suganuma, S. Shirakawa, and T. Nagao, “A genetic programming approach to designing convolutional neural network architectures,” in *Proc. Genet. Evol. Comput. Conf.*, 2017, pp. 497–504.

[37] S. Ahmed, F. Khan, A. Ghaffar, F. Hussain, and S. H. Cho, “Finger-counting-based gesture recognition within cars using impulse radar with convolutional neural network,” *Sensors*, vol. 19, no. 6, p. 1429, 2019.

**Shahzad Ahmed** received the B.S. degree from Air University, Islamabad, Pakistan, in 2012, the M.S. degree from the University of Engineering and Technology, Taxila, Pakistan, in 2016, and the Ph.D. degree in electronic engineering from Hanyang University, Seoul, South Korea, in 2021.

He is currently serving as a Postdoctoral Researcher with Radar Computing Laboratory, Hanyang University. His current research interests include radar-based gesture recognition, biomedical signal and image processing, digital healthcare, and machine-learning-based signal estimation and classification.

**Wancheol Kim** received the M.S. degree in electronic engineering from Hanyang University, Seoul, South Korea, in 2005, where he is currently pursuing the Ph.D. degree in electronic engineering.

His current research interests include radio communication system and radar signal processing.

**Junbyung Park** received the B.S. degree in computer science from Sangmyung University, Seoul, South Korea, in 2020. He is currently pursuing the Ph.D. degree in electronic engineering with Hanyang University, Seoul.

His research interests include signal processing and gesture recognition using IR-UWB and FMCW radars.

**Sung Ho Cho** (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from The University of Utah, Salt Lake City, UT, USA, in 1989.

From 1989 to 1992, he was with the Electronics and Telecommunications Research Institute, Daejon, South Korea, as a Senior Member of Technical Staff. He then joined the Department of Electronic Engineering, Hanyang University, Seoul, South Korea, in 1992, where he is currently a Full Professor, and has been the Director of the Radar Computing Laboratory since 2010. His research interests include applied signal processing, machine learning, radar computing, digital health, and smart space.