AI-BASED INDOOR LOCALIZATION USING MMWAVE MIMO CHANNELS AT 60 GHZ

Shubham Khunteta[1], Ashok Kumar Reddy Chavva (Senior Member, IEEE)[1], Avani Agrawal[1]
[1]Beyond 5G Team, Samsung R&D Institute India-Bengaluru

NOTE: Corresponding author: Shubham Khunteta, sk.khunteta@samsung.com

Abstract – *In recent years, indoor localization using wireless systems has been an important area of research for its applications towards health, security and the tracking of users. A Global Positioning System (GPS) is considered as the best solution for localization for outdoor scenarios but it fails to provide accurate positioning for indoor scenarios. Wi-Fi fingerprinting methods using received signal strength from multiple access points are popular for solving indoor localization problem. As the wireless systems move towards higher frequencies, higher bandwidth and a large antenna array, sensing has also become feasible along with communication, which is an important research area towards 6G named as Integrated Communication And Sensing (ISAC). ISAC relies on sensing parameter estimations, such as estimation of fine range, Doppler and angular information which contains the signature of the surrounding objects. A localization problem can be solved by analysing the sensing parameters. In this paper, we propose a solution for the localization problem for IEEE 802.11ay WLAN systems based on signal processing and Machine Learning (ML) in indoor scenarios. First, signal processing is used to estimate the channel in a Doppler and angular domain which separates the signal reflected from the different objects based on their range, velocity and the angular placement. Then, an ML model is used to localize the objects in the different parts of the indoor environment. We use a state-of-the-art ML algorithm such as feed forward neural networks. Further, we evaluate our algorithm for a scenario where there is a room with a transmitter and receiver, and on a dataset generated by a simulator provided by the National Institute of Science and Technology (NIST). We show that the proposed algorithm for localization, which predicts the number of persons in different parts of a room, achieves accuracy of 99% at Signal to Noise Ratio (SNR) of* 18 *dB and is able to count up to eight persons in a room with 99% accuracy at SNR greater than* 0 *dB.*

Keywords – 6G, AI, AI in wireless, integrated sensing and communication, localization, ML, mmWave, RF sensing

1. INTRODUCTION

Integrated Sensing And Communication (ISAC) is an emerging field and an important area towards 6G where along with the information transfer, sensing is also kept in focus while designing the system [1]. Sensing using the existing communication resources pave the path for ISAC. For indoor scenarios, a Wi-Fi signal is a viable option for sensing. Sensing revolves around the knowledge of the scattering of rays from the objects, which provide a unique signature about them. The reflected signal at the receiver, helps to localize the objects or even identify them. Machine learning techniques help in identifying patterns or a signature and can map it to the location in the local map [1].

With the advancement of 5G, there have been significant improvements in system frequency and bandwidth. With the increment in system frequency and bandwidth, the wavelength becomes shorter, which results in fine range estimation for object detection [2]. Radio Detection And Ranging (RADAR) has been used for detecting objects. In literature, in order to detect objects, a channel is estimated and a range-Doppler or range-angle heatmap is produced [3], [4]. A change in environment causes changes in these heatmaps. The heatmap is unique for each different environment and is considered a fingerprint of the context or abstract local map. With the uniqueness in the heatmap, localization and sensing has become an area of interest in the next generation systems [2].

Research in ISAC has been done to find the methods for coexistence of communication systems and RADAR. In [5], an architecture of 5G communication systems and RADAR is designed on chip. As 5G mmWave bands have a higher bandwidth compared to a legacy system and it is expected to be even higher for next generation systems, this allows communication systems to be a great platform for high resolution sensing. In [6], several techniques such as waveform design, sensing signal architecture, and antenna distribution are discussed for coexistence of communication and sensing systems. In [7], convergent 6G communication, localization and sensing systems are defined and its possible solutions are discussed which also involves Artificial intelligence (AI) and Reconfigurable Intelligent Surfaces (RIS). Possible designs of 6G localization and sensing systems are discussed in [2] with a 100 GHz frequency system, RIS and advance signal processing techniques. For applications related to biomedicine and security, Simultaneous Localization And Mapping (SLAM) to automatically construct maps of complex indoor environments are also discussed in the paper. In [2], challenges to make ISAC feasible, are also discussed, such as high-accuracy cm-level positioning and high-resolution 3D sensing/imaging, efficiently sharing resources in time, frequency and space domains, leveraging the real-time energy-efficient AI/ML techniques.

An Indoor Positioning System (IPS) using radio waves has been an area of interest in literature and it is seen as an alternative to Global Positioning System (GPS) in indoor scenarios. A Wi-Fi system provides better connectivity than a cellular system and is usually used for indoor localization [8][9]. Signal strength-based localization methods have been popular in literature where multiple access points are used for triangulation to estimate the location of an object of interest. These methods fail to provide fine range resolution with limited access points, performs poorly in a Non-Line-Of-Sight (NLOS) environment and provide a sub-meter level of accuracy [10].

Localization using radio signals has always been an area of interest. For an indoor scenario, Wi-Fi is viable and a better option than a cellular system. Several methods have been proposed for indoor localization using Wi-Fi signals such as [8], which describes the implementation of a Wi-Fi fingerprinting method using a Received Signal Strength Indicator (RSSI) from access points to determine the position of users in indoor areas. In [9], multiple Wi-Fi sources around the indoor area are used for localization and object tracking and the algorithm primarily relies on a triangulation method.

As the RSSI-based method suffers in NLOS conditions, in [10], location-specific Channel State Information (CSI) is used as a fingerprint and is claimed to provide 5cm precision in an 20cm×70cm area in a non-line-of-sight office environment with one link measurement. In [11], a solution for Simultaneous Localization And Mapping (SLAM) is proposed using channel state information. It captures the local spatial geometry of the area using CSI in a channel chart in such a way that points which are closer in space also appears closer in the channel chart. This is done by extracting features from the channel and applying dimensionality reduction algorithms. In [12], a RADAR is used to capture the 2-D RADAR image of the surroundings and that is used as an input for iterative closest point algorithms to solve a SLAM problem.

In this paper, we present an algorithm based on signal processing and ML for indoor localization in an ISAC kind of setup. We also propose a model for counting the total number of persons in the environment. We use an indoor scenario such as a room and use a radio signal in the form of the channel estimation field of Wi-Fi signal (IEEE 802.11ay). At the receiver, we estimate the channel in the Doppler and angular domain to extract the features that are relevant for estimation of the users' location such as range, velocity and the angular information of the users. We refer these features as sensible features further in this paper. Finally, we propose machine learning models to map the features to localize and count the persons in the surroundings. Main contributions of the paper are as follows:

- Estimating channel in angular domain for Wi-Fi signals (IEEE 802.11ay).
- Methods for extracting sensible features from the raw data of Wi-Fi signals (IEEE 802.11ay) at the receiver for localization tasks.
- Proposing a machine learning model for the localization of persons in the surrounding environment.

This paper is organized as follows: Section 2 describes the system model. In Section 3, sensible features are extracted followed by the ML model description in Section 4, and results are presented followed by the conclusion in Section 5.

2. SYSTEM MODEL

We consider an indoor scenario with a transmitter (Tx) and a receiver (Rx) which communicate using IEEE 802.11ay packets. An IEEE 802.11ay packet is of duration T_p seconds which contains preamble, Channel Estimation Field (CEF) and data symbols. Tx and Rx communicate using an antenna array of N_{tx} and N_{rx} elements respectively. Packets are transmitted with a carrier frequency of 60 GHz. In this paper, we focus on CEF for a person's localization in the indoor scenario. CEF consists of Golay Sequence of length 1024 as shown in Fig. 1 where A^{512} and B^{512} are the complementary sequences of length 512 which follow the following property:

$$\psi_{A^{512},A^{512}}(k) + \psi_{B^{512},B^{512}}(k) = 0; \forall k \neq 0, \textit{(1)}$$

where $\psi_{A^{512},A^{512}}(k)$ and $\psi_{B^{512},B^{512}}(k)$ are $k-th$ element of an autocorrelation sequence A^{512} and B^{512} respectively. Note that, in Fig. 1, the Golay sequence of length 1024 is re- peated twice and padded with smaller length complemen- tary sequences for better channel estimation in presence of inter-symbol interference and thus the CEF length is N_s. A^{512} and B^{512} also follow the following property and said to be orthogonal to each other [13]:

$$\psi_{A^{512},B^{512}}(k) = 0; \ \forall k, \qquad \textit{(2)}$$

where $\psi_{A^{512},B^{512}}(k)$ is $k-th$ element of the cross-correlation function of sequence A^{512} and B^{512}.

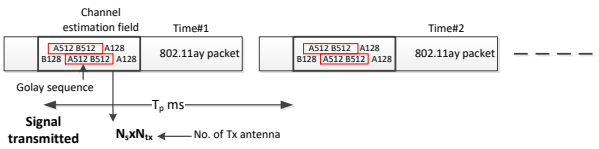

Fig. 1 – Transmitted signal structure of an 802.11ay system: Packets are transmitted with a period T_p seconds over N_{tx} antennas. The signal of interest is Channel Estimation Field (CEF) which consists of a Golay sequence of length 1024 containing two complementary sequences A^{512} and B^{512}.

Without Loss of Generality (WLOG), the surrounding environment is assumed to be static and only persons are assumed to be moving with a velocity of v m/s. The transmitted signal gets reflected from the moving persons and

environment and the resultant signal is received at N_{rx} receiver antennas. We consider an $L+1$ tap channel (H) and it can be shown as $H = [H_0, H_1, H_2, ..., H_L]$ where the channel for each tap (H_i) can be represented as [14]:

$$H_i = \begin{bmatrix} h_{1,1}(i) & \cdots & h_{1,N_{tx}}(i) \\ \vdots & \ddots & \vdots \\ h_{N_{rx},1}(i) & \cdots & h_{N_{rx},N_{tx}}(i) \end{bmatrix}, \quad (3)$$

where $h_{m,n}(i)$ is the channel coefficient for i^{th} tap between m^{th} receiver antenna and n^{th} transmitter antenna. The received signal on N_{rx} antennas can be represented as follows:

$$Y = HX + N, \quad (4)$$

where $Y = [\mathbf{y}(0), ..., \mathbf{y}(N_s + L - 1)]$, where $\mathbf{y}(k) = [y_1(k), .., y_{N_{rx}}(k)]^T$ and N is the noise vector. Transmitted signal matrix(X) is filled with shifted CEF in the rows as follows:

$$X = \begin{bmatrix} \mathbf{x}(0) & \cdots & \cdots & \mathbf{x}(N_s-1) & \mathbf{0}_{N_{tx},1} & \cdots & \mathbf{0}_{N_{tx},1} \\ \mathbf{0}_{N_{tx},1} & \mathbf{x}(0) & \cdots & \mathbf{x}(N_s-1) & \mathbf{0}_{N_{tx},1} & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \mathbf{0}_{N_{tx},1} \\ \mathbf{0}_{N_{tx},1} & \cdots & \mathbf{0}_{N_{tx},1} & \mathbf{x}(0) & \cdots & \cdots & \mathbf{x}(N_s-1) \end{bmatrix}, \quad (5)$$

where $\mathbf{x}(i) = [x_1(i), ..., x_{N_{tx}}(i)]^T$ is a vector comprising of all the i^{th} elements of the CEF sequence transmitted from all the transmitter antennas.

As shown in Fig. 2, at the receiver, the channel is estimated in the tap domain. Channel taps can be perceived as a differentiator for the arrival rays at the receiver being reflected from the persons in the surroundings based on time of arrival and thus captures the information to distinguish the persons located in different parts of the surroundings. Further, we convert the channel in the Doppler domain, which further distinguishes the persons based on their movement speed. Then, the channel is converted in an angular domain, which then separates the persons based on their angular separation. The channel in the angular domain provides the sensible features compared with the raw data at the receiver. These sensible features are then fed to a deep learning-based detector. It consists of feed-forward layers and a *softmax* layer at the output, which predicts the number of persons in all parts of the surrounding area.

3. ALGORITHM DESCRIPTION

In this section, we describe the algorithm to extract sensible features in sections 3.1-3.3 which are channel estimation and conversion to a Doppler and angular domain. And finally, deep learning-based detector is explained in Section 3.4.

3.1 Channel estimation

In this section, we describe the method for the channel estimation for IEEE 802.11ay signal which is explained in Section 2. Using *(4)*, the channel can be estimated as least square estimate [15] as follows:

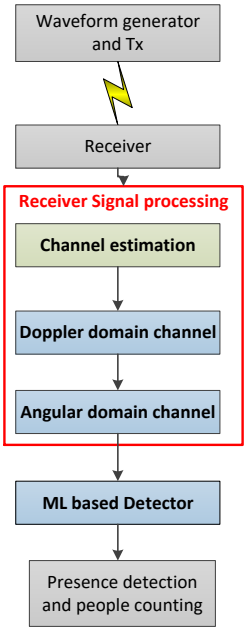

Fig. 2 – System model: Signal processing on the raw inputs followed by a machine learning model-based detector for localization

$$\widehat{H} = YX^T(XX^T)^{-1}. \quad (6)$$

Note that using *(6)*, we get an L-tap channel for every transmitter and receiver antenna pair and dimension of H is $N_{rx} \times N_{tx} \times L$ for each and every received packet. Channel taps can be seen as a discriminator for arrival rays at the receiver after reflecting off the objects, which are present in the surroundings. Channel taps depict the Time Difference of Arrival (TDoA) of reflected rays from the various objects in the surroundings, and TDoA of various rays depend on the location of the objects. There can be a scenario where reflected rays from the two objects fall under the same tap of the channel. In this scenario, the two objects cannot be distinguished solely based on the channel taps. We exploit the velocity of the objects to distinguish them further, which is explained in the next section.

3.2 Channel in delay-Doppler domain

If there is only one object that falls under a channel tap, the phasor corresponding to the channel tap rotates with a speed (let us assume ω radian per samples) that is equivalent to the Doppler or velocity of that object. Fourier Transform (FT) of the discrete time signal corresponding to this tap, results in a peak at ω in the frequency domain. Let us consider a scenario where two objects fall under the same tap of channel. If there are two objects that fall under a channel tap, then the signal corresponding to the tap is the sum of the two phasors corresponding to these two objects and the FT of this signal result in two peaks in frequency domain. The two peaks will be at the frequencies corresponding to the speed of the objects. As we know,

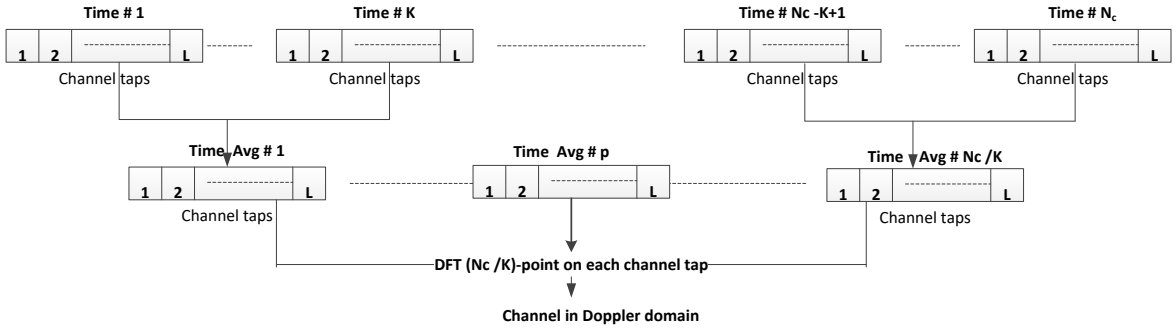

Fig. 3 – Channel conversion to Doppler domain: There are total N_c packets and channel of K consecutive packets are grouped together for averaging. An N_c/K-point DFT is performed for each channel tap.

Fourier transform resolves any signal in its constituent components, thus Fourier transform performed on a tap in the time domain will resolve the Doppler of the objects, which fall under the same tap.

We consider N_c consecutive packets in the time domain for Doppler domain channel processing. The duration of a packet is T_c seconds. The channel is estimated first for all these N_c packets as explained in Section 3.1 and can be represented as follows:

$$H^t = [H(1), ..., H(K), ..., H(N_c)]. \tag{7}$$

For coverage enhancement or to gain in the Signal to Noise Ratio (SNR), the averaging of K consecutive channel estimates is done which results in N_c/K size time domain vector of averaged channel estimates ($H^{t_{avg}}$).

$$\begin{aligned} H^{t_{avg}} =& [\text{mean}(H(1), .., H(K)), .. \\ & .., \text{mean}(H(K(\frac{N_c}{K} - 1)), .., H(N_c)]. \end{aligned} \tag{8}$$

Then N_c/K-point Discrete Fourier Transform (DFT) is performed on each channel tap of $H^{t_{avg}}$ for every tx-rx antenna pair to get the channel in the Doppler domain as follows:

$$H_l^{Dopp}(d)[m, n] = \text{DFT}(H_l^{t_{avg}}[m, n])(d), \tag{9}$$

where $l \in \{0 \cdots L\}$ is the tap index of the channel, $m \in \{1 \cdots N_{rx}\}$ and $n \in \{1 \cdots N_{tx}\}$ are the receiver and transmitter antenna index and $d \in \{1 \cdots \frac{N_c}{K}\}$ is the Doppler index or d^{th} bin of DFT. Multiple objects, which fall under the same tap index, can be distinguished now in the Doppler domain. Note that if the two objects which are in same channel tap, have a similar velocity then there is a possibility that these two objects fall under the same tap and Doppler bin and then they can't be distinguished with the Doppler domain channel.

Let us consider a scenario where there are two persons moving with velocity v_{h1} and v_{h2} respectively and they fall under the same channel tap. Phasors corresponding to both the persons rotate with the speed ω_{h1} and ω_{h2} radian per samples respectively. A sampling period of the signal $H^{t_{avg}}$ is KT_c, ω_{h1} and ω_{h2} can be represented as follows:

$$\begin{aligned} \omega_{h1} &= 2\pi F \Delta\tau_1, \\ \omega_{h2} &= 2\pi F \Delta\tau_2, \end{aligned} \tag{10}$$

where F is the carrier frequency. $\Delta\tau_1$ is the time difference of the rays which are arriving at the receiver, associated with the phase change during the sampling period and can be further simplified and written in terms of velocity as: $\frac{v_{h1} K T_c}{c}$. Similarly, $\Delta\tau_2$ can be written as $\frac{v_{h2} K T_c}{c}$. A property of Fourier Transform (FT) states that the difference between two consecutive constituent components should be larger than the inverse of number of resolvable bins. Velocity resolution (V_{res}) i.e. minimum velocity difference between two objects, which is needed for them to distinguish, can be estimated using the FT property as follows:

$$\begin{aligned} & \omega_{h2} - \omega_{h1} > \frac{2\pi}{N_c/K} \\ \implies & \frac{2\pi F K T_c (v_{h2} - v_{h1})}{c} > \frac{2\pi}{N_c/K} \end{aligned}$$

putting $F = c/\lambda$,

$$\begin{aligned} & \implies \frac{K T_c (v_{h2} - v_{h1})}{\lambda} > \frac{1}{N_c/K} \\ & \implies v_{h2} - v_{h1} > \frac{\lambda}{N_c T_c} \\ & \implies V_{res} = \frac{\lambda}{N_c T_c}. \end{aligned} \tag{11}$$

As shown in *(11)*, velocity resolution depends on the number of the samples used for Doppler domain processing and the time duration of a packet.

3.3 Channel in delay-Doppler-angular domain

Let us consider a scenario where two objects fall under the same channel tap and Doppler bin (i.e. $\Delta v < \frac{\lambda}{N_c T_c}$) as shown in *(11)*, they cannot be distinguished solely based on the Doppler domain channel, so we further explore the

spatial dimension to make them differentiable. The objective of the conversion of the Doppler domain channel to angular domain, is to resolve the transmit and arrival paths of rays into angular bins. A channel is said to be the sum of multiple paths that originate from the transmitter and arrive at the receiver as follows [16]:

$$H = \sum_i a_i e_r(\Omega_{ri}) e_t(\Omega_{ti}), \tag{12}$$

where a_i is the attenuation associated with the i^{th} path. The i^{th} path makes the angle ϕ_{ri} with the receiver antenna array and ϕ_{ti} with the transmitter antenna array and Ω_{ri} and Ω_{ti} are the respective direction cosines. $e_t(\Omega)$ and $e_r(\Omega)$ are the transmitted and received unit spatial signature, respectively, along the direction Ω and is calculated as follows [16]:

$$e_r(\Omega) = \frac{1}{\sqrt{(N_{rx})}} \begin{bmatrix} 1 \\ exp(-j2\pi\Delta_r\Omega) \\ \vdots \\ exp(-j2\pi(N_{rx}-1)\Delta_r\Omega) \end{bmatrix}, \tag{13}$$

$$e_t(\Omega) = \frac{1}{\sqrt{(N_{tx})}} \begin{bmatrix} 1 \\ exp(-j2\pi\Delta_t\Omega) \\ \vdots \\ exp(-j2\pi(N_{tx}-1)\Delta_t\Omega) \end{bmatrix}, \tag{14}$$

where Δ_r and Δ_t are the separation between consecutive antennas normalized by wavelength in the receiver and transmitter antenna arrays. For received signal space, an orthonormal basis can be written as follows [16]:

$$\sigma_r = \{e_r(0), e_r(\frac{1}{L_r}), ..., e_r(\frac{N_{rx}-1}{L_r})\}, \tag{15}$$

where L_r is length of the receiver antenna array normalized by wavelength. Similarly, the basis for transmit signal space can also be constructed as σ_t. Receive and transmit signals can be represented in the angular domain using basis σ_r and σ_t. Transmitted signal (X) and received signal (Y) can be represented in the angular domain as follows [16]:

$$\begin{aligned} X^a &= U_t^* X, \\ Y^a &= U_r^* Y, \end{aligned} \tag{16}$$

where U_t and U_r are the unitary matrices in the signal spaces σ_t and σ_r respectively and can be calculated as follows [16]:

$$\begin{aligned} U_t(k,l) &= \frac{1}{\sqrt{N_{tx}}} exp(\frac{-j2\pi kl}{N_{tx}}), \\ U_r(k,l) &= \frac{1}{\sqrt{N_{rx}}} exp(\frac{-j2\pi kl}{N_{rx}}), \end{aligned} \tag{17}$$

where $k, l \in \{0 \cdots N_{tx} - 1\}$ for U_t and $k, l \in \{0 \cdots N_{rx} - 1\}$ for U_r. The angular domain channel can be calculated by putting *(16)* into *(4)* as follows:

$$H^a = U_r^* H^{Dopp} U_t. \tag{18}$$

The dimension of H^a are $N_{rx} \times N_{tx} \times L \times \frac{N_c}{K}$, where the first two dimensions are for angular bins for receiver and transmitter, third dimension is for channel taps and the last dimension is for Doppler bins. This angular domain channel helps in distinguishing the objects further based on angular separation and is used as the input for a deep learning model.

3.4 Deep learning model for localization

In this section, the architecture of a Deep Learning (DL) model is described. WLOG, let us assume the surrounding environment is divided into multiple sectors (a total of N sectors). There are a maximum of X persons present in the surroundings with a maximum of M persons in each sector. The DL model predicts how many persons are present in each sector, which is referred to as a localization model. We train another DL model, which is referred to as a counting model and it is trained to predict the total number of persons which are present in the surrounding environment (a maximum of X persons). Absolute value of the channel in the angular domain is used as the input.

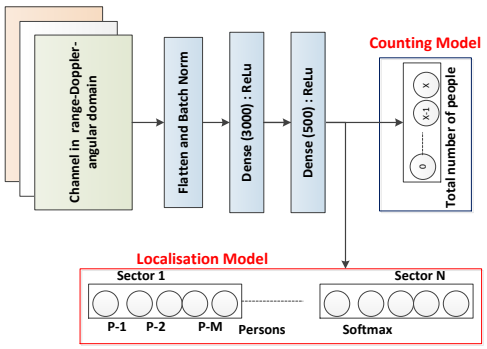

Fig. 4 – Deep learning model: Channel in range Doppler angular domain is the input and at output layer, two models are proposed. A counting model which predict the count, the total number of persons in the surroundings (maximum is X) and a localization model which predicts the number of persons in each area of surroundings (a total of N sectors and a maximum of M persons in each sector). Note that, we are using two different neural networks with the same structure for localization and counting activities.

As shown in Fig. 4, absolute values of the angular channel is fed to a couple of $dense$ layers and followed by a *softmax* layer. For the localization model, the output layer consists of M *softmax* neurons for each of the N sectors and for the counting model, the output layer is a *softmax* layer of X neurons. The counting model is used for post-processing of the prediction of the localization model for improving the accuracy.

Fig. 5 describes the algorithm for post-processing of the localization model prediction based on the counting model prediction. Let us consider a scenario where the counting model predicts the total number of persons in the environment is P_c. The localization model predicts the total number of persons for each sector as $\{P_L^1, \cdots P_L^N\}$, so the total number of persons in the surroundings predicted by the localization model is $P_L = \sum_i^N P_L^i$. If P_c and P_L are different, then the results of the localization model are up-

dated based on the counting model prediction and this is because, we found during the experiment that the counting model provides better accuracy than the localization model in all the scenarios. The localization model provides the *softmax* probabilities for the number of persons present in each sector. We sort the sectors based on the difference between *softmax* probabilities of the top-2 neurons. In Fig. 5, $B1_{si}$ is the best neuron and $B2_{si}$ is the second best neuron for a sector i and A is a list of sorted sectors based on the difference between *softmax* probabilities of the $B1_{si}$ and $B2_{si}$). Until the P_L matches with the P_c, sectors are chosen from the top of the sorted array A and the prediction of the localization model is updated with the second best *softmax* neuron if it helps in minimizing the gap between P_L and P_c . Updates are stopped when P_L matches with the P_c.

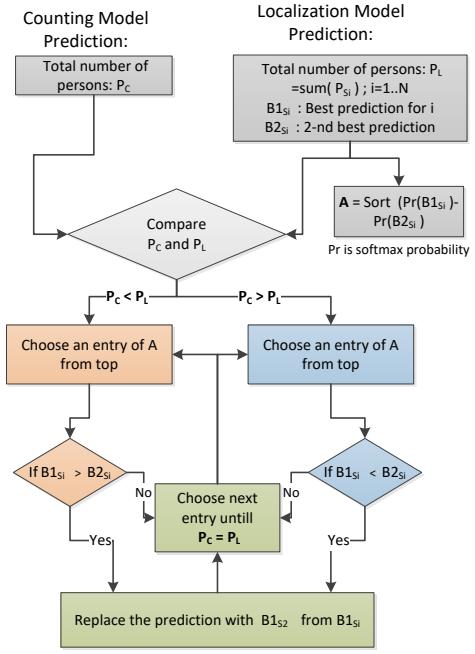

Fig. 5 – Post-processing for the localization model predictions: Total number of persons predicted by the counting model is P_C and by the localization model is P_L. Result of the localization model is updated with this algorithm until P_L matches with P_c.

4. EXPERIMENTAL SETUP AND PERFORMANCE EVALUATION

For the experimental setup, we have considered a dataset provided by National Institute of Technology and Science, USA. The dataset is generated using an IEEE 802.11Aay WLAN simulator [17],[18] and it considers an indoor scenario, specifically a room of dimension (7.8m ×7m ×3m), which is divided into a total of 9 sectors (sector A to I) as shown in Fig. 6. There are two access points which communicate using IEEE 802.11ay packets, one of which is the transmitter and located at $(-3.9, 0, 2.8)$ and the other is the receiver located at $(3.9, 0, 2.8)$. There can be a maximum of four persons in any sector and maximum eight persons in the room. In this simulator, persons are modelled as 17 joints for scattering of rays and velocity of persons is around ≈ 1 m/s. The aim is to find the number of persons in each sector using the received signal. For example, in Fig. 6, there is one person each in sector G and C and the rest of the sectors have 0 persons.

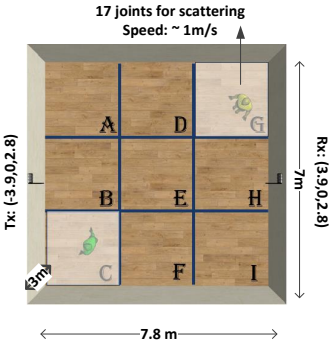

Fig. 6 – Experimental setup: A room is considered for dataset generation and it is divided into nine sectors. Samples are generated with different arrangements of the persons in the room. For example, in this figure, there is a person in sector C and G and the rest of the sectors are empty.

In this dataset, a training sample corresponds to the transmission of an IEEE 802.11ay packet repeating $N_c = 128$ times using $N_{tx} = 4$ antennas. We take only the CEF part of the packets for sensing or localization. The dimension of the signal transmitted is $2432 \times 4 \times 128$ which is $(N_s \times N_{tx} \times N_c)$. The carrier frequency is 60 GHz, sampling frequency is 1.76 GHz and the number of channel taps are $L = 45$. At the receiver, the signal is received at $N_{rx} = 4$ antennas and its dimension is $2476 \times 4 \times 128$ $(N_s + L - 1 \times N_{rx} \times N_c)$.

For training of the DL model, a total of 15578 different arrangements of persons in the room are considered. Data is generated for multiple Signal to Noise Ratio (SNR) points ranging from -18dB to 18dB. In the DL model, there are two dense layers of 3000 and 500 neurons respectively. We have used the Rectified Linear unit ($ReLu$) for activation with L2 regularization. For the localization model, 5 $softmax$ neurons are used for each sectors summing up to a total of 45 neurons. For the counting model, total 9 $softmax$ neurons are used. To avoid overfitting, we have used an ADAM optimizer with learning rate 0.0005 and exponential decay every 10000 steps where 1 step training involves 32 samples and the model is trained for 200 epochs. The dataset is imbalanced and there are more samples for less number of persons in any sector, so a weighted learning has been used for different labels. For testing, samples are generated with different noise vectors using the simulator.

The channel is estimated as mentioned in Section 3.1 and its dimensions are $4 \times 4 \times 45 \times 128$, where the 4^{th} dimension represents the number of packets (N_c) and the other three represent the same information as mentioned in Section 3.1. Then the channel is averaged for $K = 8$ consecutive times and processed for Doppler domain conversion (Section 3.2) and the dimensions of this channel come out to

be $4 \times 4 \times 45 \times 16$. Then, they are converted to the angular domain (Section 3.3) and absolute value of this channel is then fed to the DL model (Section 3.4).

We have drawn the results for both counting model and localization model. For performance evaluation, accuracy of the counting model is drawn and shown in Table 1 which is number of times the total number of persons in the room are predicted correctly. As shown in Table 1, for the dataset with good signal condition, the model predicts correctly for almost all the samples.

Table 1 – Counting model accuracy: % Number of samples for which the model predicts correctly the total number of persons in the room.

SNR (dB)	Accuracy%
18	99.90
10	99.82
5	99.76
0	99.60
−5	97.23
−10	74.34
−18	39.01

We have also drawn the counting model results for different numbers of persons present in the room. In the dataset, there are upto eight persons present in the room. We plot the accuracy of the counting model w.r.t number of persons (total 8 plots correspond to the number of persons present in the room) for various SNR points and is shown in Fig. 7.

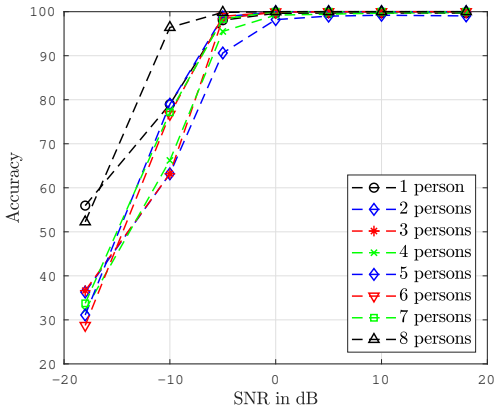

Fig. 7 – Counting model accuracy plot w.r.t different number of persons in the room vs SNR.

The localization model is also evaluated and its accuracy drawn and shown in Table 2 which is the number of times the total number of persons in each sector is predicted correctly. Here, even if prediction of one sector mismatches from the true label out of nine sectors, we mark it a failure case for the sample. Only if the prediction of all the sectors matches with the true labels, the sample is marked as a success case. Hence, the accuracy of the localization model is less than the counting model. For poor signal conditions, such as SNR -18dB, there is not enough signal information for the DL model to make an accurate prediction.

As we explained earlier that if the localization model fails to predict only one sector out of the total nine sectors,

Table 2 – Localization model accuracy: % Number of samples in which model predicts correctly the number of persons in each sector.

SNR (dB)	Accuracy%
18	99.00
10	96.09
5	93.68
0	83.13
−5	44.74
−10	5.97
−18	2.00

we mark that sample as a failure for localization model accuracy. In Fig. 8, we compare the accuracy of sector-specific prediction for different numbers of persons with SNR. Here, we consider the prediction of each sector as an independent prediction and show that the accuracy of sector-wise prediction degrades with the increment in the number of the persons in the sectors. At an SNR greater than 0 dB, accuracy of the sector-wise prediction is greater than 95% for any number of persons in the sector. At a lower SNR i .e. SNR lower than −10 dB, accuracy of the sector-wise prediction decreases with the increment in the number of persons as shown in Fig. 8.

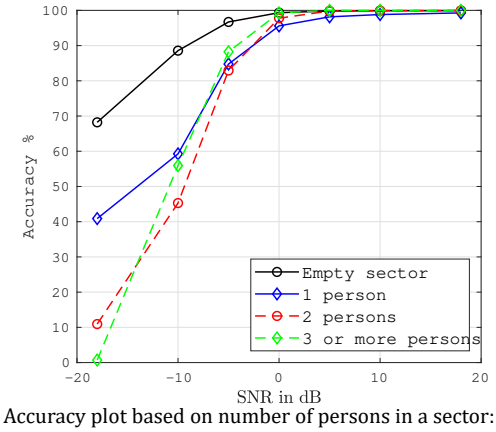

Fig. 8 – Accuracy plot based on number of persons in a sector: Localization model prediction accuracy when the sectors are considered as independent samples and the plot shows the accuracy vs SNR performance with different numbers of persons in the sectors.

Finally, we discuss the resolutions in the range, velocity and angular domain. In this paper, first we have tried to distinguish the persons as far as possible using channel estimation and signal processing. We have used machine learning to analyse the extracted features for localization. Two persons can be distinguished if they fall under different taps which means delay in arrival of rays scattered from the two persons should be greater than 0.56 ns (=1/1.76GHz). This information suggests if the separation of persons is more than 17 cm, they can be distinguished. Similarly, using *(11)*, velocity resolution is 4 cm/sec where λ is 5mm as the carrier frequency is 60GHz, N_c is 128 and T_c is 1ms. Considering 16 Doppler domain bins, two persons with a velocity difference up to 64 cm/sec can be distinguished if they are at least 4 cm/sec apart in velocity. For angular resolution [16], it can be calculated using a normalized length of antenna array (L_r and L_t) which is 2 with N_{rx} and N_{tx} being 4 and resolvable angular bins

are of size $1/L_r$ and $1/L_t$ for the receiver and transmitter respectively i.e. 0.5 and 0.5 radian which is 28^o for both transmitter and receiver.

5. CONCLUSION

In this paper, we proposed an indoor localization method, which can assist the research of Integrated Sensing And Communication (ISAC). We observed near-perfect localization accuracy with the collected data. The proposed algorithm contributes to the emerging ISAC technology and is easy to implement. We used both signal processing and machine learning to separate out multiple persons and localize them, which is a novel method. Our future work involves person identification along with localization and contribution towards simultaneous location and mapping.

ACKNOWLEDGMENT

We thank the International Telecommunication Union (ITU) AI/ML 5G challenge team and NIST, USA for providing the dataset and simulator for generating the dataset. The problem statement was one of the open challenges of ITU AI/ML challenge in 5G, 2021 and it was organised by NIST, USA. The solution provided in this paper won the challenge for problem statement PS002 in the 2021 edition of ITU AI/ML for 5G challenge.

REFERENCES

[1] Danny Kai Pin Tan, Jia He, Yanchun Li, Alireza Bayesteh, Yan Chen, Peiying Zhu, and Wen Tong. "Integrated Sensing and Communication in 6G: Motivations, Use Cases, Requirements, Challenges and Future Directions". In: (2021). DOI: 10.1109/JCS52304.2021.9376324.

[2] Andre Bourdoux. "6G White Paper on Localization and Sensing". In: *arXiv:2006.01779v1 [eess.SY]* (2020).

[3] X. Gao et. al. "RAMP-CNN: A Novel Neural Network for Enhanced Automotive Radar Object Recognition". In: *arXiv:2011.08981v2 [eess.SP]* (2022).

[4] X. Gao et. al. "Experiments with mmWave Automotive Radar Test-bed". In: *arXiv:1912.12566v3 [eess.SP]* (2022).

[5] H. Aghasi and Heydari P. "Millimeter-Wave Radars-on-Chip Enabling Next-Generation Cyberphysical Infrastructures". In: *IEEE Communications Magazine, DOI: 10.1109/mcom.001.2000544* (2021).

[6] T. Wild et al. "Joint Design of Communication and Sensing for Beyond 5G and 6G Systems". In: *IEEE Access, Volume: 9, DOI: 10.1109/ACCESS.2021.3059488* (2021).

[7] Carlos D Lima et. al. "Convergent Communication, Sensing and Localization in 6G Systems: An Overview of Technologies, Opportunities, and Challenges". In: *IEEE access, DOI 10.1109/ACCESS.2021.3053486* (2021).

[8] H. Chabbar and M. Chami. "Indoor localization using Wi-Fi method based on Fingerprinting Technique". In: *International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), pp. 1-5, doi: 10.1109/WITS.2017.7934613* (2017).

[9] Noor A. K. Z. and Muayad S. C et. al. "Indoor Localization System Using Wi-Fi Technology". In: *Iraqi Journal of Computers, Communications, Control and System Engineering (IJCCCE), Vol. 19, No., DOI: 10.33103/uot.ijccce.19.2.8* (2019).

[10] C. Chen and Y. Chen et. al. "High accuracy indoor localization: A WiFi-based approach". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6245-6249, doi: 10.1109/ICASSP.2016.7472878* (2016).

[11] C. Studer and S. Medjkouh et. al. "Channel Charting: Locating Users within the Radio Environment using Channel State Information". In: *arXiv:1807.05247* (2018).

[12] J. W. Marck et. al. "Indoor radar SLAM A radar application for vision and GPS denied environments". In: *European Radar Conference, pp. 471-474.* (2013).

[13] Tseng and Liu. "Complementary set of Sequences". In: *IEEE Transactions on Information Theory, Vol. IT-18, No. 5* ().

[14] S. Wang and Abdi. "A. Aperiodic Complementary Sets of Sequences-Based MIMO Frequency Selective Channel Estimation". In: *IEEE Communication Letters, Vol. 9, No. 10* ().

[15] E. G. Larsson and P. Stoica. "Space-Time Block Coding for Wireless Communications". In: *Cambridge, UK: Cambridge University Press* (2003).

[16] Tse and Vishwanath. "Fundamentals of Wireless Communication". In: *Chapter: 7.3.4 Angular Domain Representation of MIMO Channels* (2005).

[17] "A collection of open-source tools to simulate IEEE 802.11ad/ay WLAN networks in network simulator ns-3". In: (2021). URL: https://github.com/wigig-tools.

[18] "Q-D simulation & Modeling framework for sensing". In: (2021). URL: https://mentor.ieee.org/802.11/dcn/21/11-21-0746-01-00bf-q-d-simulation-modeling-framework-for-sensing.pptx.

AUTHORS

Shubham Khunteta received a Bachelor of Technology degree in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 2014. He joined Samsung R&D Institute India Bangalore, India in July 2014, where he currently is an engineer in a research and development team that works on differentiating solutions for mobile devices. His research interests include algorithm design for the physical layer, beam management, machine learning-assisted communications, integrated sensing and communication, 5G NR systems, 6G, millimeter-wave and machine learning for communications.

Ashok Kumar Reddy Chavva (M'06-SM'14) received his Bachelor of Technology degree in electronics and communications engineering from the Jawaharlal Nehru Technological University, Hyderabad, India, in 2003, and an M.E. degree in telecommunication engineering from the Indian Institute of Science, Bangalore, India, in 2005. In June 2005, he joined a wireless startup, Beceem Communications, which later became part of Broadcom. Here, he was involved in developing physical layer algorithms for the first 4G system based on WiMAX and LTE. He worked with Broadcom till November 2013. Since November 2013, he has been with Samsung R&D Institute India Bangalore, India, where he currently leads an R&D team that works on beyond-5G system design. He is currently pursuing his Ph.D. degree in electrical communication engineering from the Indian Institute of Science. His research interests include algorithm design for the physical layer, performance evaluation of wireless communication systems, 5G NR systems, 6G, millimeter-wave and tera-hertz systems, and machine learning for communications. He received the best paper award at IEEE CCNC, Las Vegas, USA, 2016 and best paper (third) at the IEEE World 5G Forum, 2020.

Avani Agrawal received a Bachelor of Technology degree in electronics and communications engineering from the International Institute of Information and Technology, Hyderabad, India, in 2020. Since June 2019, she has been with Samsung R&D Institute India Bangalore, India and is a member of the Beyond 5G team in the Mobile Communications Department. Her interest areas include machine learning, computer vision and algorithm development for software.