Solids on Soli: Millimetre-Wave Radar Sensing through Materials

Klen Čopič Pucihar, Nuwan Attygalle, Matjaz Kljun, Christian Sandor, Luis Leiva

Solids on Soli: Millimetre-Wave Radar Sensing through Materials

KLEN ČOPIČ PUCIHAR, University of Primorska, Slovenia
NUWAN T. ATTYGALLE, University of Primorska, Slovenia
MATJAŽ KLJUN, University of Primorska, Slovenia
CHRISTIAN SANDOR, School of Creative Media, City University of Hong Kong, Hong Kong SAR
LUIS A. LEIVA, University of Luxembourg, Luxembourg

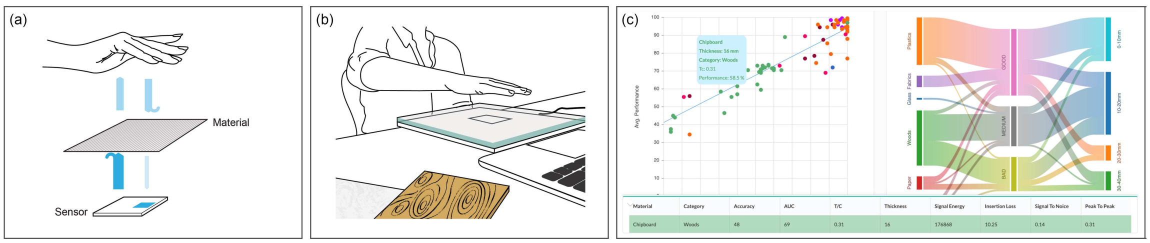


Fig. 1. Integrating radar-based gesture recognition in consumer devices requires the signal to pass through a covering material twice (a). To help application designers select suitable materials for interaction, we provide an ample catalogue of 75 everyday materials that is agnostic to the underlying gesture classifier. Designers can evaluate their own recognisers against just 3 reference materials (b), enter the observed performance measures into a simple tool (c) and get updated performance estimates for all materials in the catalogue.

Gesture recognition with miniaturised radar sensors has received increasing attention as a novel interaction medium. The practical use of radar technology, however, often requires sensing through materials. Yet, it is still not well understood how the internal structure of materials impacts recognition performance. To tackle this challenge, we collected a large dataset of 14,090 radar recordings for 6 paradigmatic gesture classes sensed through a variety of everyday materials, performed by humans (6 materials) and a robot system (75 materials). Next, we developed a hybrid CNN+LSTM deep learning model and derived a robust indirect method to measure signal distortions, which we used to compile a comprehensive catalogue of materials for radar-based interaction. Among other findings, our experiments show that it is possible to estimate how different materials would affect gesture recognition performance of arbitrary classifiers by selecting just 3 reference materials. Our catalogue, software, models, data collection platform, and labeled datasets are publicly available.

CCS Concepts: • **Hardware** → *Sensor devices and platforms*; • **Human-centered computing** → **Gestural input**; *Interaction design process and methods.*

Additional Key Words and Phrases: Radar Interaction; Materials; Soli; Gestures; Deep Learning

Authors' addresses: Klen Čopič Pucihar, klen.copic@famnit.upr.si, University of Primorska, Slovenia; Nuwan T. Attygalle, nuwan.attygalle@famnit.upr.si, University of Primorska, Slovenia; Matjaž Kljun, matjaz.kljun@famnit.upr.si, University of Primorska, Slovenia; Christian Sandor, christian@sandor.com, School of Creative Media, City University of Hong Kong, Hong Kong SAR; Luis A. Leiva, name.surname@uni.lu, University of Luxembourg, Luxembourg.

Manuscript submitted to ACM

ACM Reference Format:
Klen Čopič Pucihar, Nuwan T. Attygalle, Matjaž Kljun, Christian Sandor, and Luis A. Leiva. 2022. Solids on Soli: Millimetre-Wave Radar Sensing through Materials. *Proc. ACM Hum.-Comput. Interact.* , (2022), 19 pages.

1 INTRODUCTION

In recent years, gesture recognition with miniaturised radar sensing has received increased attention in academia and industry. Two factors have fuelled this trend: the emergence of low-cost low-power radar chips and the impressive breakthroughs in deep learning, which allows to interpret radar signals very accurately for interaction. However, there are still several unsolved research problems in order to enable practical applications of this technology.

First and foremost, real-world interaction with radar devices often requires sensing through materials. This not only includes interaction e.g. with a mobile phone while in our pocket, but also interaction with other devices (e.g. infotainment systems or smart homes) through a radar sensor that is integrated in nearby objects such as clothing accessories, car dashboards, seats, doors, indoor furniture, and walls. Nevertheless, it is still not well understood how different materials affect gesture recognition performance. Ideally, application designers would speed up development with a convenient method for determining which materials would perform best for a recogniser of their choice.

To make radar-based gesture recognition practical on different materials, we first need to understand how the radar signal degrades when passing through them. This seemingly straightforward problem is rather challenging in reality. To measure signal degradation with high precision, one would need to acquire an expensive vector network analyser and operate it in a shielded environment, preventing electromagnetic interference. Another problem arises from the plethora of potential materials and various thicknesses one would like to evaluate. Together with different gesture sets and custom classifiers, this would result in a vast number of measurements to perform. Lastly, if we were able to overcome all these aforementioned problems, how could we conveniently predict gesture recognition performance on untested materials when only signal characteristics (e.g. transmission coefficient) for the materials are known?

To address these challenges we (i) used Google Soli, a millimetre-wave (mm-wave) radar sensor, to record range Doppler images and time series data for 11 core features as well as 9 meta-features; (ii) developed a state-of-the-art hybrid CNN+LSTM deep learning model to test recognition performance on 75 materials; and (iii) derived a robust method to measure signal distortions, which we used to compile a catalogue of materials, and predict recognition performance on arbitrary gesture classifiers. Even though a few tabulations of material properties are available in the literature [16, 34], to the best of our knowledge none exist for a large variety of everyday materials in the mm-wave band.

2 PROBLEM STATEMENT AND CONTRIBUTIONS

Radar technology uses radio-frequency (RF) electromagnetic waves to detect nearby objects. Essentially, a radar device has a transmitter antenna (Tx) that emits an RF pulse to the environment, and a receiving antenna (Rx) that captures the echoed pulse. The received signal is analysed to determine object properties, such as their radar cross-section and velocity [51].

RF waves can be affected by a variety of phenomena such as: absorption, refraction, diffraction, polarisation, scattering, and reflection [45]. Since these phenomena happen simultaneously and interact with one another, a simplified model based on the lightwave analogy (Figure 2) is used to characterise the signal, where only the incident, reflected, and transmitted components are considered. This model is typically employed in network analysis, where designers and

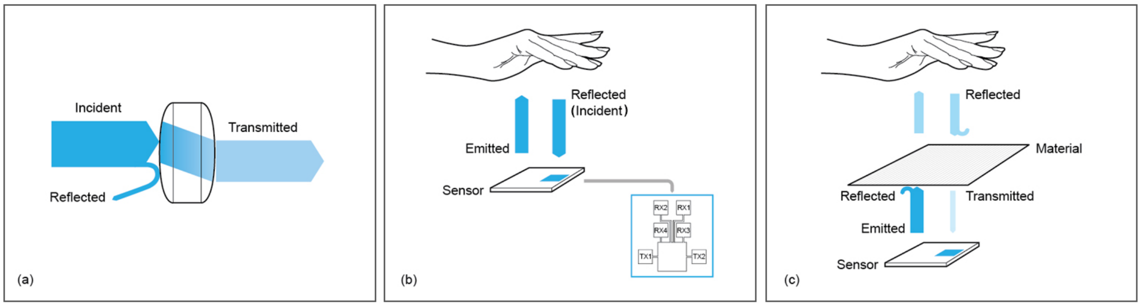


Fig. 2. Lightwave analogy (a) to characterise signal distortions [3]. When no occluding material is present (b), the transmitted RF pulse gets reflected by the user's hand (referred to as *incident signal*). When an occluding material is present (c), the reflected signal needs to pass though the material (referred to as *transmitted signal*).

manufactures of networking components characterise how such components distort the input signal over a desired frequency range [3].

When radar sensing is used for interaction through materials, the RF signal needs to pass through them twice. Since different materials have different characteristics, which depend on their thickness and their dielectric properties, these characteristics will significantly affect how the signal propagates through them. This opens up the following research questions:

(1) How to accurately characterise radar signal distortions through materials without using an expensive equipment?
(2) How would such distortions affect gesture recognition?
(3) Can we predict performance of arbitrary classifiers when only the characteristics of signal distortions for a given material are known?

While it may appear obvious that the contrast of an RF signal degrades while it passes through materials, thereby affecting recognition accuracy, our work is the first one to precisely quantify this degradation. As previously discussed, such a quantification is non-trivial yet critical when designing radar-based gesture interaction systems.

Among other findings, we observed a strong inverse correlation of signal amplitude and material thickness for several Soli core features that we describe in Section 3.5. We validate our findings on different material types, showing that our proposed measurement method is adequate and reveals that some materials are more apt for interaction than others. As a result, we have compiled an extensive catalogue of everyday materials for radar sensing that we make publicly available (see Supplementary Materials).

3 RELATED WORK

We analyse previous research according to our main areas of interest: gesture interaction, RF sensing technologies, mm-wave radar sensors (including Google Soli), and sensing through materials.

3.1 Gesture interaction

Gesture interaction is an active research topic with a history dating back to the 1960s with Sutherland's Sketchpad project [55] and his far reaching vision of the Ultimate Display essay [56]. Today, gesture interaction can be categorised into 2 broad groups: (i) *mid-air*, used for example in consumer electronics such as gaming consoles, and (ii) *stroke-based*, used for example in devices with touchscreens such as smartphones. We focus on the former group, given the increasing importance it has gained recently.

Mid-air gesture interaction has been extensively researched as an alternative to other modes or as a complementary mode of interaction in a variety of settings such as entertainment [6, 44, 47], automotive industry [19, 28, 39, 46, 50], medical applications [10, 26, 31, 43, 54], wearable computing [4, 14, 15, 17], smart home control [18, 60, 64], virtual reality manipulation [23, 65], and art installations [36, 37]. Such interaction is particularly interesting where other modes are dangerous, hard, or impossible to use.

3.2 RF sensing technologies

Despite popular technologies used for implementing gesture recognisers such as RGB [27, 52, 58] or infrared (IR) [11, 15, 21, 53, 54] cameras, RF-based solutions including radar [29, 40], Wi-Fi [2, 41, 69], GSM [68], and RFID [14] offer several advantages. Above all, RF sensing technologies are insensitive to light, which usually affects camera and, especially, IR based solutions (both cannot be used in bright sunlight). RF sensing does not require an elaborate setup of various sensors on or around users. In addition, the RF signal can penetrate non-metallic surfaces and can sense objects and their movements through them.

RF sensing has been used for analysing walking patterns or gait [7, 33, 62], tracking sleep quality and breathing patterns [43, 70], and recognising movements of body parts such as hands for interactive purposes [14, 25, 28, 36, 37, 41, 60]. The radars used in these studies operated at various frequencies, ranging from 2.4 GHz [60, 70] to 24 GHz [28, 43].

3.3 Millimeter-wave radar-on-chip sensors

To detect and recognise fine-grained interactions, it is necessary to increase the radar's spatial resolution. For this, radar chips working at even higher frequencies, around 50–70 GHz, have been recently used [29, 63]. Such chips open up the path to precise close-range gesture interactions in a variety of applications, including wearable, mobile, and ubiquitous computing. Since these sensors operate in the millimeter range, they allow for tighter integration of the circuit due to the reduced size of different passive (non-moving) components and low-power requirements [29]. These properties also allow manufacturing them inexpensively at scale.

Radar sensing is very effective in detecting close-proximity, subtle, nonrigid motion mostly articulated with hands and fingers (e.g. rubbing, pinching, or swiping) [22, 61] or with small objects (e.g. pens) [63]. It can also recognise large gestures in 3D space [35] with remarkable accuracy. Recent research has explored radar-based interaction with everyday objects and in augmented reality scenarios [9, 59], as well as creating music [5, 48]. A mm-wave radar can also distinguish various materials when placed on top of it [25, 66]. What is missing, however, is an investigation of gesture recognition performance *through* various materials present on and around us, which is the focus of our work.

We should note that there are essentially two standard approaches to mm-wave gesture recognition: one feeds the raw signals or derived images (e.g. Doppler images) directly into a classifier [20, 22], while the other approach applies different beamforming vectors to extract/track location before feeding it to a classifier [35, 42]. The results of the characterisation in this work are primarily focused on the first approach, since we use the Google Soli sensor. The second approach is possible with other mm-wave sensors like the IWR1443 board from Texas Instruments, for example.

3.4 Sensing through materials

The fact that RF signals can penetrate non-metallic materials makes them particularly interesting for interaction. Alas, sensing through materials has been barely explored. Notable examples include tracking people through walls [1, 8] and gesture recognition through walls [41] and above an office desk [25]. All these approaches have focused on coarse gestures instead of fine-grained ones, have considered only one material, and have not used radar-on-chip sensors.

Leiva et al. [20] investigated radar-based gesture recognition on wearable devices, but they did not characterise signal distortions nor estimated recognition performance on arbitrary gesture classifiers. Now that mm-wave radar technology is available on consumer products, it is expected that it will be further integrated in a variety of objects in the near future.

3.5 Google Soli

Soli is a 60 GHz 4-channel receiver (Rx antennas) 2-channel transmitter (Tx antennas) radar-on-chip that has become available in consumer electronics such as the Pixel 4 smartphone. Complemented with time-varying micro-Doppler frequency features analysis [7], the sensor offers detection of movements with near-millimeter precision [29]. Soli comes with an SDK that can represent the radar signal with range Doppler images as well as a variety of low-level *core* features [22]:

(F1) *Range:* Overall distance of the moving targets to the sensor.
(F2) *Acceleration:* Overall acceleration of the moving targets.
(F3) *Energy total:* Amount of reflected energy overall.
(F4) *Energy moving:* Amount of reflected energy from the moving targets.
(F5) *Velocity:* Overall velocity of the moving targets.
(F6) *Velocity dispersion:* Dispersion of energy over the Doppler space.
(F7) *Spatial dispersion:* Dispersion of energy over the range space.
(F8) *Energy strongest component:* Amount of reflected energy from the most dominant moving target.
(F9) *Movement index:* Moving target identifier.
(F10) *Fine displacement:* Instantaneous velocity of each moving target.
(F11) *Velocity centroid:* Weighted average of the overall velocity.

These core features essentially characterise the energy distribution across the radar transformation space, which have been shown to accurately describe relative finger dynamics [5, 22] as well as end-effector trajectories [28, 67].

4 GESTURE RECOGNITION EXPERIMENTS

We describe the gesture recogniser we developed under a control condition (no occluding material). It will be our baseline classifier to analyse radar sensing performance through different materials later on.

4.1 Gesture set

We surveyed previous work that used mm-wave radar sensing for interaction [61] as well as gestures supported by the Soli sensor in consumer devices. From these gesture sets, we chose the ones that are relatively easy for a mechanical system to replicate. Eventually, a set of 6 distinct gestures was selected (Figure 3).

4.2 Experimental system

To explore the effect of different materials on gesture classification performance, we should be able to execute as many times as needed the gestures with minimal articulation variation. For this reason, we first built a robot system and then recruited human participants.

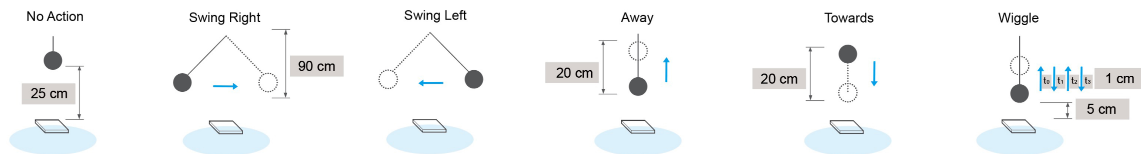


Fig. 3. Experimental gesture set. Filled and dashed circles denote, respectively, the initial and final position of the ball.

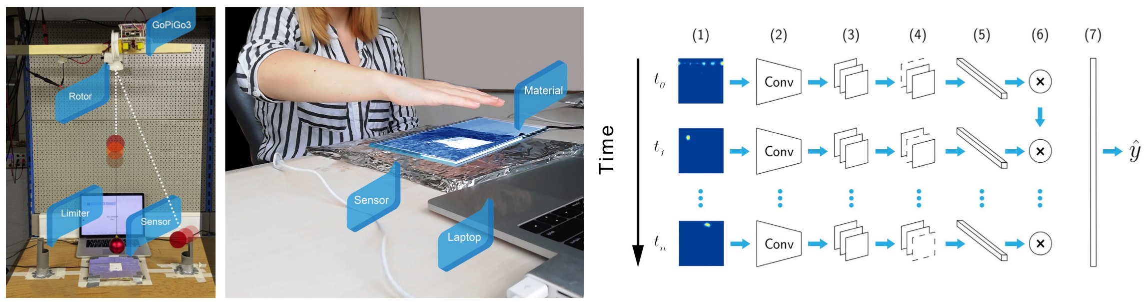


Fig. 4. Experimental platform setup and model arhitecture. Left: robot setup. Middle: human setup, Right: Hybrid deep learning model architecture. Range Doppler images (1) are processed with a CNN (2) that extracts feature maps followed by max pooling (3) and spatial dropout (4) layers. Then, a fully connected layer (5) creates the feature vectors for a recurrent LSTM layer with dropout (6) and finally a softmax layer (7) outputs the gesture class prediction ($\hat{y}$).

4.2.1 Robot system. We followed the setup proposed by Leiva et al. [20]. It consists of (i) a robot-based gesture simulator based on the GoPiGo3, to which an empty plastic ball of 5 cm diameter is attached; (ii) an aluminium-shielded frame placed on the table, on which different materials can be placed (the frame prevents the radar signal from escaping around the analysed material); and (iii) the Soli radar sensor placed in the frame and connected to a computer (see Figure 4 left). The distance between the sensor and the ball in its lowest position is 5 cm.

This system is designed to generate two types of movements: pendulum-like (e.g. swing or swipe movements) and vertical movements along the z-axis. The pendulum movement is generated by manually releasing the ball from a limiter position, whereas the vertical movements are automatically generated by the robot.

4.2.2 Human system. This setup is identical to the robot system, but instead of a bouncing ball, a human hand executes the gestures (Figure 4 right). The user is sitting in front of a computer display that instructs what gesture to perform and when to perform it. Besides visual indicators, the system also plays a sound to help the user executing the gesture systematically within a fixed time window. For this setup, we recruited six participants (5 males, 1 female) aged 24–34.

4.3 Data collection setup

Gestures are stored in two data formats: (i) a sequence of frames of 32×32 px range Doppler images; and (ii) a time series of 11 core features, computed with the Soli SDK. We use the range Doppler images for developing our classifiers, since they enable robust recognition [20, 61]. The extracted core features are used to characterise signal distortions.

The Soli sensor is configured to record at 1000 Hz, Doppler range sensitivity is set to [-2,0] dB (based on pilot trials with our robot system), and the built-in adaptive clutter filter is disabled. Note that 1000 Hz sampling represents an upper bound, which can be conveniently downsampled as needed.

We collected 20 repetitions of the 6 gestures for 75 materials with the robot system (9,000 recordings) and 200 trials of each gesture in the baseline condition (1,200 recordings). Each participant performed 10 repetitions of the 6 gestures for 6 materials (2,160 recordings), and 300 trails (50 per participant) for each gesture in the baseline condition, totalling 1,800 recordings. Some recordings were flagged as inappropriate by the experimenter (e.g. the user articulated a different gesture or it was performed sloppily) and thus were removed, reducing the total number of baseline recordings performed by the participants to 1,730.

Since Soli computes range Doppler images for each Rx antenna, we averaged them to ensure a robust frame representation. Further, images were grayscaled and sequences were resampled to 100 Hz and padded to 400 timesteps, which is large enough to accommodate for arbitrary gesture articulations. As a reference, each recorded gesture took 1.5 seconds on average, or 150 timesteps.

4.4 Model architecture

Our model, depicted in Figure 4, is a hybrid deep CNN+LSTM (convolutional neural network + long short-term memory) model, inspired by previous work [12, 22, 30, 32, 61]. We also tested a Conv3D architecture (see Appendix) but it did not match the excellent performance of the hybrid CNN+LSTM architecture. See Table 5 for a comparison. Note that by having such an excellent baseline performance, we can attribute the subsequent recognition performance degradation to the occluding materials. Otherwise, the recogniser would become a confounder variable in our experiments.

Each frame (Doppler image) is processed by a stack of $32 \times 64 \times 128$ convolutional layers with 3×3 filters to capture spatial information. The resulting frame sequence is further processed in a recurrent fashion by means of an LSTM layer (embedding size of 128) to capture temporal information, and eventually classified with a softmax layer. Our model has 2.4M weights, which is rather small for today's standards.

Each convolutional layer automatically extracts feature maps from input frames that are further processed by max pooling and spatial dropout layers. The max pooling layers (pool size of 2) downsample the feature maps by taking the largest value of the map patches, resulting in a local translation invariance.

Crucially, the spatial dropout layer (drop rate of 0.25) removes entire feature maps at random, instead of individual neurons (as it happens in regular dropout layers), which promotes independence between feature maps, thus improving performance. The LSTM layer uses both a dropout rate and a recurrent dropout rate of 0.25. The softmax layer has dimensionality of 6, since we have 6 gestures.

4.5 Model training and evaluation

We created random splits comprising 50% of the data for model training, 20% for model validation, and the remaining 30% for model testing. The test data are held out as a separate partition, which simulates unseen data. The model was trained only with data from 'no material' condition in batches of 10 sequences using categorical cross-entropy as loss function.

We used the Adam optimiser with learning rate $\eta = 0.0005$ and decay rates $\beta_1 = 0.9, \beta_2 = 0.999$. The maximum number of epochs was set to 200, but we also set an early stopping criteria of 50 epochs. That is, training stopped if the validation loss did not improve after 50 consecutive epochs, and the best model weights were retained.

We built 7 classifiers for each data source (robot and humans), totaling 14 classifiers. As shown in Table 1, all classifiers were built for different gesture combinations (6/6 and all possible combinations of 5/6 gestures), to demonstrate that our findings are agnostic to the recogniser and data source. Finally, each classifier is evaluated on the baseline (no material) condition as well as on 75 materials for the robot system and on 6 materials for the human system, respectively.

4.6 Results

Table 1 shows that our model architecture performs really well for different combinations of gestures: accuracy is 98% on average for the robot and around 96% for human participants. G0 denotes no action, which is used as a "rejection class". Further evaluation on all materials are provided in Table 5 in the Appendix.

Performance is slightly better in the robot condition (Table 1 top), as expected, since the variation in gesture articulation is lower as compared to that of the humans, as the robot movements are systematically executed the same way.

Table 1. Gesture recognition performance metrics for the baseline (no material) condition, using both our robot system (top table) and aggregated data from six users (bottom table).

Classifier	Gesture set	Recordings	ACC	AUC	Precision	Recall	F1
Robot A	[G0 G1 G2 G3 G4 G5]	1,200	97.35	98.42	97.40	97.35	97.35
Robot B	[G1 G2 G3 G4 G5]	1,000	96.07	97.57	96.20	96.07	96.10
Robot C	[G0 G1 G2 G3 G4]	1,000	99.29	99.55	99.31	99.29	99.29
Robot D	[G0 G1 G2 G3 G5]	1,000	96.55	97.88	96.75	96.55	96.56
Robot E	[G0 G1 G2 G4 G5]	1,000	98.97	99.35	98.97	98.97	98.97
Robot F	[G0 G1 G3 G4 G5]	1,000	98.57	99.11	98.60	98.57	98.57
Robot G	[G0 G2 G3 G4 G5]	1,000	99.29	99.55	99.29	99.29	99.29
		Mean	98.01	98.78	98.07	98.01	98.02
Humans A	[G0 G1 G2 G3 G4 G5]	1,730	93.33	96.00	94.05	93.33	93.30
Humans B	[G1 G2 G3 G4 G5]	1,428	92.36	95.25	92.58	92.36	92.38
Humans C	[G0 G1 G2 G3 G4]	1,440	97.00	98.12	97.02	97.00	96.99
Humans D	[G0 G1 G2 G3 G5]	1,451	97.22	98.27	97.41	97.22	97.22
Humans E	[G0 G1 G2 G4 G5]	1,450	98.96	99.34	98.96	98.96	98.96
Humans F	[G0 G1 G3 G4 G5]	1,451	95.33	97.08	95.37	95.33	95.34
Humans G	[G0 G2 G3 G4 G5]	1,440	95.83	97.38	96.32	95.83	95.80
		Mean	95.72	97.35	95.96	95.72	95.71

Gestures:
G0 - No action
G1 - Swing right
G2 - Swing left
G3 - Away
G4 - Towards
G5 - Wiggle

ACC: Accuracy
AUC: Area Under the ROC
F1: F1 score

5 MODELING SIGNAL DISTORTION

To characterise distortion of electromagnetic waves as they pass though materials, we need to measure the difference between incident and transmitted signal across the frequency range of the radar sensor. As discussed in Section 1, one approach is to use a vector network analyser, which is an expensive equipment that needs to be operated in an isolated environment. As an alternative, we propose to perform indirect measurements using only information from the radar sensor and our robot system. The incident signal can be measured when no material is present, whereas the transmitted signal can be measured when different materials occlude the sensor. Therefore, we can measure distortions in the Amplitude and Signal-to-Noise Ratio (SNR). We formulated the following hypotheses:

- H1: Distortions in amplitude of the radar signal should be detectable using indirect measurements of core features.
- H2: Distortions in SNR should be detectable using indirect measurements also available through core features.

5.1 Data collection

We recorded the radar signal passing through 7 different materials (*Oriented strand board*, *Paper*, *Drywall*, *Acrylic*, *Polycarbonate*, *Polyethylene*, and *Styrofoam*) each of various thicknesses (26 materials in total) as our robot performed 2 pendulum swings over the sensor. We chose this particular gesture as it induces a signal that is sinusoidal in nature with a clear amplitude, period, and phase, which should be easy to notice visually. To make our observations robust to measurements error, we repeated the recording 20 times for each material (520 recordings in total).

5.2 Data analysis

5.2.1 Detecting amplitude distortions with indirect measures (H1). Since we completed multiple recordings of the pendulum swing gesture for each material, we aggregated the data prior to visualisation. Then, we synchronised the first peak or valley on the same time series. In sum, our aggregation procedure followed these steps:

(1) Downsample the recording frame rate from 1000 to 200 Hz, which is more than enough to capture fine-grained variations of Soli core features.
(2) Apply Dynamic Time Warping with barycenter averaging.
(3) Smooth the time series by calculating the mean over a sliding window of 10 frames.
(4) Apply Savitzky-Golay finite impulse response (FIR) filter with polynomial order 3 and frame length of 21.

We opted for a barycentric averaging method because the arithmetic mean depends on the order of frame aggregation, which can be problematic when tying to create reliable descriptors [38]. Further, we selected a relatively small window size and small polynomial order to avoid aggressive filtering [49].

We note that core feature F10 (Fine Displacement) requires additional preprocessing to compensate for measurement drifts caused by the sensor, therefore we performed 2 additional steps:

(1) Outlier removal using a Hampel filter with a window size of 100 frames and removal criteria of 3 standard deviations.
(2) Offset the signal to zero using the minimum value in the whole time series.

With this setup, we should observe a drop in signal amplitude with increasing material thickness. We also should observe a periodic behaviour of the extracted features (e.g. distinct peaks at a constant period). To reinforce the visual observation of amplitude and material thickness, we ran a correlation analysis between peak-to-peak amplitude of time series data between core features and the thickness of a given material.

5.2.2 Detecting distortions in SNR (H2). Similar to signal amplitude, we expect to observe a distortion in SNR when the signal passes through various materials. We also expect to see an inverse correlation betwewn SNR and material thickness (the greater the thickness, the smaller the SNR).

SNR is defined as the ratio of the power of a signal to the power of background noise. We performed an indirect computation of SNR, as follows. First, instead of considering signal power we considered signal amplitude. Second, the amplitude measured in the “no action” gesture cases (Figure 3), in which the ball is stationary above the sensor, is considered background noise. To get a reliable estimation, we took 20 measurements from each of the above-mentioned 26 materials, and aggregated these measurements by averaging at each frame of each time series. We also performed frame outlier removal with the interquartile range technique [57] to ensure consistent estimates.

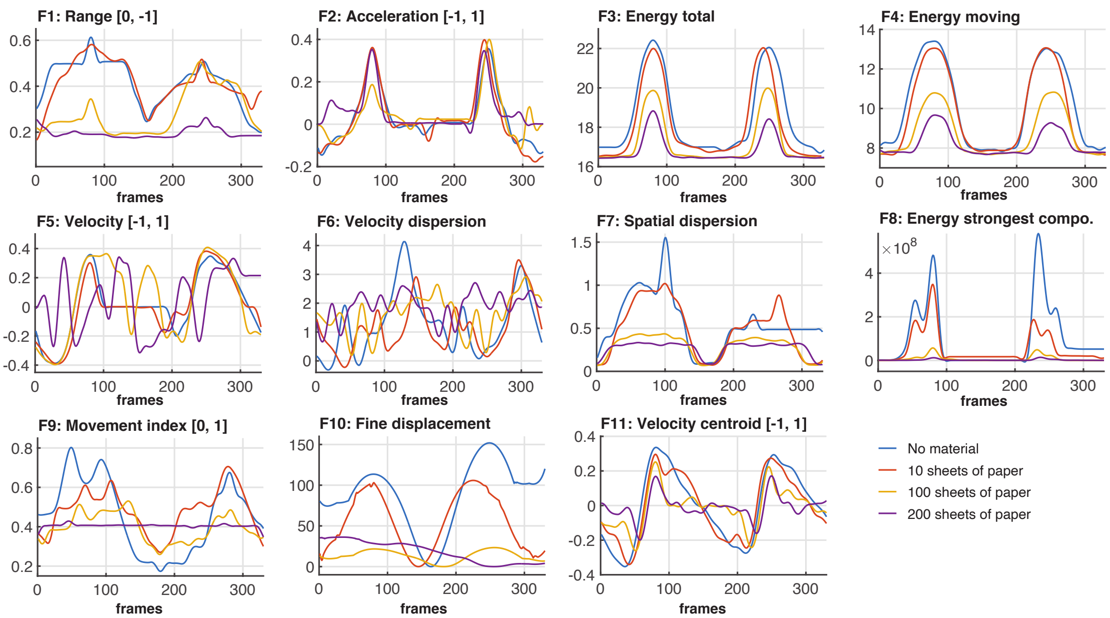


Fig. 5. Time series of the radar signal as the ball swings past the sensor. Each plot represents one Soli core feature for the baseline (no material) condition and 3 different thicknesses of *Paper* material (10, 100, and 200 sheets of paper). See our Supplementary Materials for additional results.

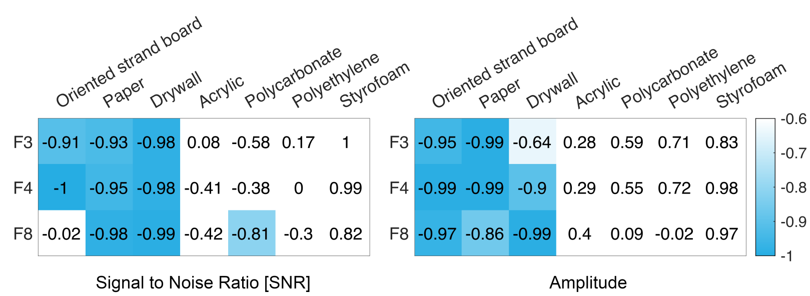


Fig. 6. Correlation between *peak-to-peak amplitude* and *Signal-to-Noise Ratio* against material *thickness* for Soli core features F3, F4, and F8.

5.3 Results

As an example, the visualisation of time series data for *Paper* materials (10, 100, and 200 sheets of paper) is provided in Figure 5. We can see that peak-to-peak amplitude drops with increased material thickness for core features *Energy total* (F3), *Energy moving* (F4), and *Energy strongest component* (F8). These features follow the periodic behaviour of a pendulum swing. On the other hand, the time series data for more transparent materials such as *Styrofoam*, *Polyethylene*, or *Acrylic*, only follow periodic swing behaviour whereas a drop in amplitude was not observed (see Supplementary Materials).

The correlation analysis of Amplitude and SNR with material thickness (Figure 6) confirmed our research hypotheses H1 and H2. Except for more transparent materials to the radar signal, the correlation is very strong overall (mean -0.92 for amplitude and mean -0.82 for SNR). We can conclude that, to select appropriate material candidates, we can use Soli core features F3, F4, and F8, since their signal distortions correlate best with material thickness.

6 MODELING MATERIAL PERFORMANCE

As the quality of the transmitted radar signal decreases due to signal distortions caused by occluding materials, we expect to see a performance drop in our gesture classifiers. We hypothesise:

- H3: Distortions of the signal amplitude and SNR caused by materials should result in a drop of recognition performance.
- H4: Performance drop can be modeled by indirect measurements of material properties based on the incident and transmitted signal.

6.1 Material properties

We define (physical) material properties as a set of meta-features that can be derived from Soli core features. Considering the nature of our experimentation system (e.g. our method cannot measure phase shift) we chose *Transmission coefficient* (T_c) and *Insertion loss* (L) to measure material properties. Following standard measures in time series analysis, we chose these descriptive statistics as meta-features: Mean, Median, Maximum, Asum (absolute sum across all values in the time series), and Energy (sum of squares across the whole time series).

Transmission coefficient is defined as the transmitted voltage divided by the incident voltage. If the absolute value is larger than 1, a system is said to have gain; otherwise it has attenuation [3]. We make the assumption that occluding materials can only induce insertion loss, hence it is reasonable to limit their T_c to 1, which we will refer to as T_c'. As our system cannot directly measure voltage, we use signal Amplitude to calculate T_c. *Insertion loss* (in dB) is a pseudo-feature based on the transmission coefficient [3]: $L = -20 \log_{10} |T_c|$. Only values larger than 0 are physically possible, which we will refer to this limited insertion loss as L'.

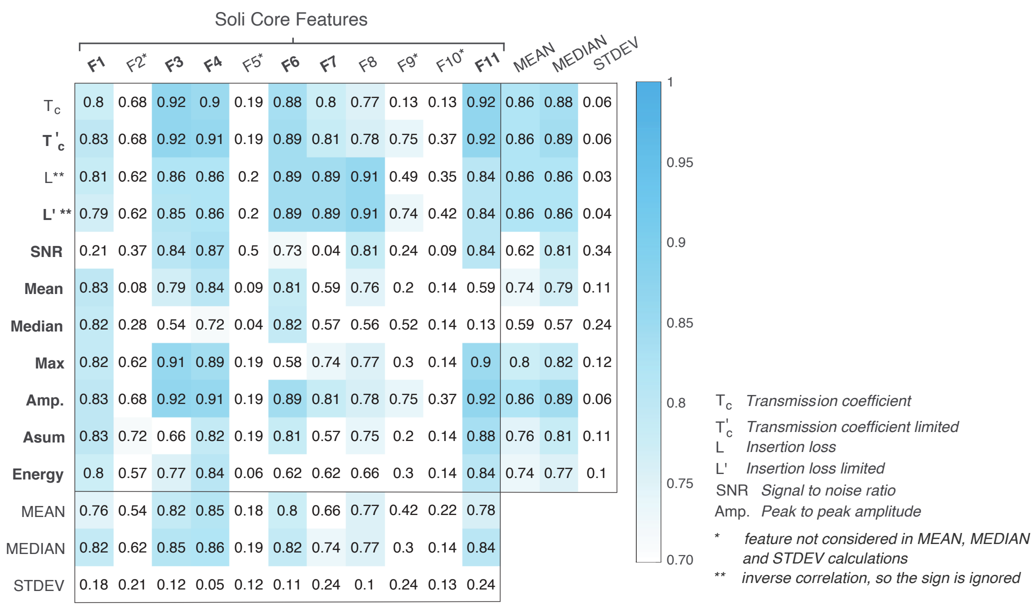
6.2 Modeling performance drop

We build linear regression models of performance drop, where performance is defined as Perf = (Accuracy + AUC)/2 to get a single prediction outcome. The procedure follows these steps:

(1) Determine which core features are good candidates for the task at hand, by conducting a correlation analysis of average performance against each material's meta-features.
(2) Fit a linear regression model of recognition performance given the statistically significant predictors (model coefficients) of material properties. We repeat this step for all the 14 gesture classifiers (Section 4.5).
(3) Choose the strongest predictor (with the highest p-value) and fit a simple linear regression model. We also repeat this step for all the 14 gesture classifiers.

6.3 Results

6.3.1 Correlation analysis for all core features. Figure 7 shows that the best overall correlating core feature is Energy moving (F4). The results also indicate that Acceleration (F2), Velocity (F5), Movement index (F9), and Fine displacement (F10) are not good predictors. For these core features, the absolute mean and absolute median values across all material meta-features are either low or have a high standard deviation; see the 3 bottom rows in Figure 7. Based on these results, we conclude that core features F1, F3, F4, F6, F7, F8, and F11 should be used for building multi-predictor linear regression model. The high correlation between performance, amplitude, and SNR (rows 5 and 9 in Figure 7) confirms H3: Distortions of the amplitude and SNR caused by an occluding material will result in a substantial degradation of recognition performance.



	F1	F2*	F3	F4	F5*	F6	F7	F8	F9*	F10*	F11	MEAN	MEDIAN	STDEV
T_c	0.8	0.68	0.92	0.9	0.19	0.88	0.8	0.77	0.13	0.13	0.92	0.86	0.88	0.06
T'_c	0.83	0.68	0.92	0.91	0.19	0.89	0.81	0.78	0.75	0.37	0.92	0.86	0.89	0.06
L**	0.81	0.62	0.86	0.86	0.2	0.89	0.89	0.91	0.49	0.35	0.84	0.86	0.86	0.03
L' **	0.79	0.62	0.85	0.86	0.2	0.89	0.89	0.91	0.74	0.42	0.84	0.86	0.86	0.04
SNR	0.21	0.37	0.84	0.87	0.5	0.73	0.04	0.81	0.24	0.09	0.84	0.62	0.81	0.34
Mean	0.83	0.08	0.79	0.84	0.09	0.81	0.59	0.76	0.2	0.14	0.59	0.74	0.79	0.11
Median	0.82	0.28	0.54	0.72	0.04	0.82	0.57	0.56	0.52	0.14	0.13	0.59	0.57	0.24
Max	0.82	0.62	0.91	0.89	0.19	0.58	0.74	0.77	0.3	0.14	0.9	0.8	0.82	0.12
Amp.	0.83	0.68	0.92	0.91	0.19	0.89	0.81	0.78	0.75	0.37	0.92	0.86	0.89	0.06
Asum	0.83	0.72	0.66	0.82	0.19	0.81	0.57	0.75	0.2	0.14	0.88	0.76	0.81	0.11
Energy	0.8	0.57	0.77	0.84	0.06	0.62	0.62	0.66	0.3	0.14	0.84	0.74	0.77	0.1
MEAN	0.76	0.54	0.82	0.85	0.18	0.8	0.66	0.77	0.42	0.22	0.78			
MEDIAN	0.82	0.62	0.85	0.86	0.19	0.82	0.74	0.77	0.3	0.14	0.84			
STDEV	0.18	0.21	0.12	0.05	0.12	0.11	0.24	0.1	0.24	0.13	0.24			

Fig. 7. Correlation analysis between gesture recognition performance and material meta-features of Soli core features. Features F2, F5, F9, and F10 are weakly correlated and so they were excluded from MEAN, MEDIAN, and STDEV row-wise calculations.

The results also show that the best correlating meta-features are Transmission coefficient and Insertion loss, where the two can be further enhanced if capped to 1 and 0, respectively. This capping of values makes sense because, as previously hinted, they cannot induce gain when measured with an occluding material. For this reason, T_c' and L' should be considered for building multi-predictor linear regression models.

6.3.2 Linear regression models. We chose all high-correlation core features (F1, F3, F4, F6, F7, F8, and F11), however we noticed that the *Energy strongest component* (F8) is linearly dependent to other core features. Thus, we excluded it and built a multi-predictor linear regression model with 54 predictors (9 material properties for each selected core feature, marked as bold on the left margin in Figure 7). This linear regression model exhibits an excellent fit (Adj. $R^2 = 0.94$, Figure 8 left). However, we can build a simpler regression model using only the most relevant material's meta-feature — T_c' for core feature *Energy Total* (F3) — and still get a very good fit (Adj. $R^2 = 0.85$, Figure 8 right).

Table 2 shows how well linear regression models can predict performance of our 14 gesture classifiers. Single and multi-predictor regression models are built in the same way as in Figure 8. The results show a good fit for all models, with an excellent Adj. R^2 for the multi-predictor model (mean=0.91, std=0.03) and a high Adj. R^2 for the single-predictor model (mean=0.81 with std=0.03 for robot and mean=0.85 with std=0.09 for human data). This confirms H4: it is possible to build a linear regression model of gesture detection performance drop using material properties that are based

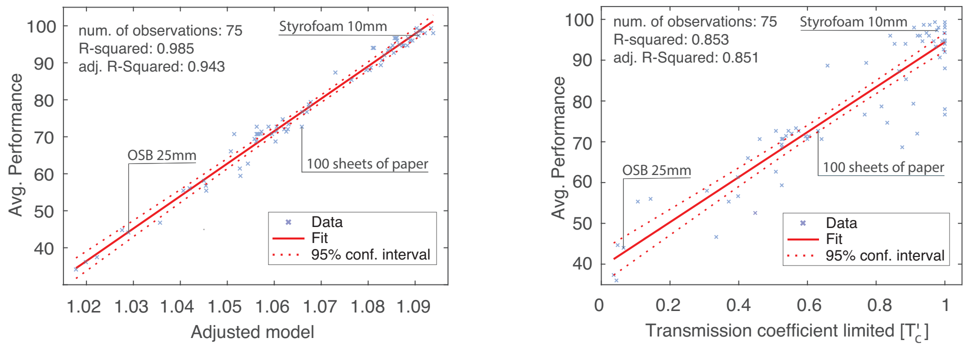


Fig. 8. Linear regression models to predict gesture classification performance. Left: multi-predictor model using 54 high-correlation Soli core features and material meta-features, marked in bold in Figure 7. Right: single-predictor model using T_c' of core feature *Energy Total* (F3).

on indirect measurement of incident and transmitted signal. This was validated for classifiers trained on different combinations of gestures. Further, to show that recognition performance can also be estimated with other model architectures, we repeated the same set of experiments with a Conv3D architecture (see Table 4 in Appendix).

Table 2. Linear regression models (LRMs) to predict performance of Hybrid architecture for different gesture classifiers trained on both robot and human data with different gestures sets. The results show strong correlation across all conditions.

	Single LRM Perf ~ Trans. coef. limited (F3)							Multiple LRM Perf ~ 54 coefficients						
Classifier (Table 1)	Materials	DoF	RMSE	MAE	R²	Adj. R²	p-value	Materials	DoF	RMSE	MAE	R²	Adj.R²	p-value
Robot A	75	73	6.78	5.09	0.85	0.85	<.001	75	20	4.20	1.64	0.99	0.94	<.001
Robot B	75	73	6.49	4.73	0.82	0.81	<.001	75	20	4.90	1.92	0.97	0.89	<.001
Robot C	75	73	8.63	6.70	0.83	0.83	<.001	75	20	4.88	1.85	0.99	0.94	<.001
Robot D	75	73	8.18	6.39	0.77	0.77	<.001	75	20	4.60	1.78	0.98	0.93	<.001
Robot E	75	73	9.30	7.48	0.79	0.79	<.001	75	20	5.74	2.21	0.98	0.92	<.001
Robot F	75	73	6.06	4.59	0.78	0.78	<.001	75	20	4.68	1.87	0.97	0.87	<.001
Robot G	75	73	6.63	4.91	0.85	0.85	<.001	75	20	6.05	2.39	0.97	0.87	<.001
Mean			7.44	5.70	0.81	0.81				5.01	1.95	0.98	0.91	

	Single LRM Perf ~ Trans. coef. limited (F3)						
Classifier (Table 1)	Materials	DoF	RMSE	MAE	R²	Adj.R²	p-value
Humans A	6	4	4.04	3.20	0.85	0.82	<.01
Humans B	6	4	3.15	2.40	0.93	0.91	<.01
Humans C	6	4	4.50	3.39	0.91	0.89	<.01
Humans D	6	4	4.46	3.19	0.85	0.81	<.01
Humans E	6	4	3.31	2.49	0.93	0.91	<.01
Humans F	6	4	4.93	3.83	0.75	0.69	<.05
Humans H	6	4	2.43	1.73	0.94	0.92	<.01
Mean			3.83	2.89	0.88	0.85	

Acronyms: *Perf: (ACC + AUC)/2 DoF: Degrees of Freedom RMSE: Room Mean Square Error MAE: Mean Absolute Error R²: R-squared Adj. R²: Adjusted R²*

7 PERFORMANCE ON ARBITRARY CLASSIFIERS

The main goal of our catalogue is to conveniently predict performance of a gesture classifier when only characteristics of signal distortions for a given material are known. This would enable designers to make more informed decisions when deciding on their own gesture sets, model architectures, and operational conditions for their sensing systems.

We evaluate how well our catalogue of materials can support the above-mentioned claim for situations where performance (measured by the geometric average of Accuracy and AUC, see Section 6.2) of a given gesture classifier is known for at least 3 materials from our catalogue. For simplicity, we use only one predictor (transmission coefficient); however, we have shown that it is possible to improve the model fit by considering more predictors (Section 6.2). At the same time, adding more predictors implies additional effort since more data points are required to build the model, hence a small number of predictors is preferable.

Note that 3 is the minimum number of materials to derive reasonable results, in order to cover the whole range of our catalogue: one material should come from the low opacity range ($T_c' \in [0.8, 1]$), one from the middle opacity range ($T_c' \in [0.3, 0.5]$), and one from the high opacity range ($T_c' \in [0, 0.2]$). We fitted the linear regression model with this single coefficient as in Section 6.3.2 and repeated this process 4 times, each time with a different combination of 3 materials.

Table 3 reports the results of these experiments. Material combinations were randomly chosen, following the procedure described above. The results show Adj. R^2 ranging from 0.75 to 0.81 and mean absolute error (MAE) ranging from 5.36% and 6.57%, which suggest that it is possible to estimate how different materials would affect gesture recognition performance by considering just 3 reference materials. Designers can thus predict gesture recognition performance for all the materials in the catalog for their own classifiers, provided that their set of gestures is similar to the one we have investigated. In the next section we elaborate more on this discussion.

We offer an interactive tool in our software repository that provides a convenient way to build various linear regression models and visualise the list of material recommendations based on predicted performance, as highlighted in Figure 1c. Our catalogue is designed as a guide book, is color coded, and has cards that include factual data about material properties together with microscope images at various magnifications (1x, 50x, 200x). We also provide in our software repository a simple step-by-step guide to expand the catalogue with new materials.

Table 3. Single-predictor linear regression models using T_c' of Energy total (F3) and observations from 3 materials to predict gesture recognition performance on all other materials.

MODEL FIT												MODEL EVALUATION					
Material 1		Material 2		Material 3													
id	name (thickness mm)	id	name (thickness mm)	id	name (thickness mm)	Num materials	DoF	RMSE	MAE	R^2	Adj.R^2	Num materials	DoF	RMSE	MAE	R^2	Adj.R^2
14	Eva Fome (40)	51	Deb Cerovy.(10)	82	Picea Abies (17)	3	1	0.96	0.52	1.00	1.00	72	72	7.26	5.82	0.81	0.81
42	Silk (<1)	60	Wood Populus (10)	32	Chipboard (32)	3	1	0.25	0.13	1.00	1.00	72	72	8.43	6.11	0.75	0.75
10	MDF (10)	21	Ceramic tiles (6)	75	Dywall (24)	3	1	9.83	5.20	0.91	0.82	72	72	7.59	5.36	0.80	0.80
35	Paper (10 sheets)	79	Paper (200 sheets)	80	Paper (300 sheets)	3	1	1.79	0.96	1.00	0.99	72	72	8.04	6.57	0.79	0.78

Gesture classifier:
Gesture set: [G0 G1 G2 G3 G4 G5]
Dataset: Robot
Material: none
Recordings: 1,200

Acronyms: *DoF: Degrees of Freedom RMSE: Room Mean Square Error MAE: Mean Absolute Error R^2: R-squared Adj. R^2: Adjusted R-squared*

8 DISCUSSION, LIMITATIONS, AND FUTURE WORK

Our work can inform application designers interested in estimating a-priori performance of materials for mid-air gesture interaction. The results for *Oriented strand board*, *Paper*, and *Drywall* (Section 5) provide evidence for the validity of our proposed approach, showing a high correlation between material thickness and both peak-to-peak amplitude and SNR (Figure 5). We can conclude that our proposed indirect measurement method is suitable for describing signal distortions as the radar signal passes though various materials. However, our results also indicate that the proposed method fails to accurately describe signal distortions for materials that are transparent to the radar signal, such as *Styrofoam*, *Polyethylene*, *Acrylic*, and *Polycarbonate*. One confounding factor can be attributed to the fact that we tested only thicknesses up to 5 cm for these materials, for which small changes in amplitude and SNR are difficult to detect by our experimental setup. This could be addressed by testing more thicknesses, although the operational range of mm-wave radar is bound to close proximity (up to 30 cm).

To make our evaluation agnostic to the data source, recogniser, and architecture, we tested up to 28 different classifiers (Table 2 and Table 4). The results show a good fit in all cases, which allows us to conclude that our catalogue can be used for predicting gesture recognition performance for new materials and similar gesture sets.

As the ultimate test, we have shown how to conveniently estimate gesture recognition performance when only characteristics of signal distortions for a few materials are known. This is possible to achieve by fitting a simple linear regression model based on performance estimates of just 3 materials from our catalogue. We argue that further improvements are possible if more reference materials are considered, but at the same time we acknowledge that it would take additional effort to do so. Overall, our companion tool provides a reasonable ballpark in a matter of minutes, if not seconds.

For the curious reader, our linear regression models to predict recognition performance have an analytical form. For example, $\text{Perf} = 42.3 + 52.7\, T_c'$ for the simple model, where T_c' is the limited transmission coefficient of the material. The intercept and slope values were averaged over the 7 robot-based classifiers considered (Table 1) across all the materials from our catalogue. This expression can be used to provide a rough estimate without having to use our companion tool, provided that the designer is using a similar classifier to the ones we have developed in this work. In the Supplementary Materials we provide the analytical expression of the full regression model, which has 54 coefficients and provides more accurate estimates than the simple model.

Our catalogue currently holds 75 materials of various types and thicknesses, but there are many more materials one could envision for radar sensing applications. Due to the low complexity of our experimental setup, one could easily expand the existing catalogue. In principle, this is limited to those users who have access to the Soli sensor and the SDK. However, it is important to note that this does not preclude the usage of our catalogue, since the extracted Soli core features are hardware-agnostic [22]. As such, it should be possible to predict gesture recognition performance for untested materials with other radar-on-chip sensors that operate in the same frequency range as Soli. We believe this would make for an interesting future work.

Another avenue for future work would be to use a vector network analyser and extend our catalogue with direct measurements of material properties. Researchers with access to such equipment could further analyse the influence of signal distortion on gesture recognition performance using our datasets. Perhaps most interesting would be to investigate distortions in signal phase, which is not possible to measure with our method. Signal phase is important for modeling distortion in electronics [3] and it may be useful to predict gesture recognition performance as well.

9 CONCLUSION

We have studied how different materials occluding a mm-wave radar sensor would affect gesture recognition performance. Our proposed method is suitable for understanding signal degradation as it passes through various materials. Critically, our method uses indirect measurements of radar signal properties, requiring only a radar-on-chip sensor and optionally a DIY robot system for automating gesture articulation. As a result, we have compiled a catalogue of everyday materials that can support designers to determine which materials and gesture sets will perform best in their particular situation. Finally, it is possible to predict gesture recognition performance on any material similar to the ones we have analysed. Our catalogue is diverse enough to cover a reasonably large range of materials, and we hope others will find it useful and build upon our work.

ACKNOWLEDGMENTS

We thank Birte Malz for helping us with the data collection, Hongbo Fu for reviewing and early draft of this paper, Google Inc. for donating a Soli sensor, and Pui Chung Wong, Jiaming Liao, Dávid Maruscsák, Fonny Phan, and Ran Ju for their help with graphics production.

REFERENCES

[1] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. 2015. Capturing the Human Figure through a Wall. *ACM Trans. Graphics* 34, 6 (2015).

[2] Fadel Adib, Zach Kabelac, Dina Katabi, and Robert C Miller. 2014. 3D tracking via body radio reflections. In *Proc. NSDI*. 317–329.

[3] Agilent 1997. Understanding the Fundamental Principles of Vector Network Analysis. Agilent AN 1287-1.

[4] Shaikh Shawon Arefin Shimon, Courtney Lutton, Zichun Xu, Sarah Morrison-Smith, Christina Boucher, and Jaime Ruiz. 2016. Exploring non-touchscreen gestures for smartwatches. In *Proc. CHI*. 3822–3833.

[5] Francisco Bernardo, Nicholas Arner, and Paul Batchelor. 2017. O soli mio: exploring millimeter wave radar for musical interaction. In *Proc. NIME*. 283–286.

[6] Amit Bleiweiss, Dagan Eshar, Gershom Kutliroff, Alon Lerner, Yinon Oshrat, and Yaron Yanai. 2010. Enhanced interactive gaming by blending full-body tracking and gesture animation. In *ACM SIGGRAPH ASIA 2010 Sketches*. 1–2.

[7] Victor C Chen, Fayin Li, S-S Ho, and Harry Wechsler. 2006. Micro-Doppler effect in radar: phenomenon, model, and simulation study. *IEEE Trans. Aerosp. Electron. Syst* 42, 1 (2006), 2–21.

[8] Kevin Chetty, Graeme E Smith, and Karl Woodbridge. 2011. Through-the-wall sensing of personnel using passive bistatic wifi radar at standoff distances. *IEEE Trans. Geosci. Remote Sens.* 50, 4 (2011), 1218–1226.

[9] Barrett Ens, Aaron Quigley, Hui-Shyong Yeo, Pourang Irani, Thammathip Piumsomboon, and Mark Billinghurst. 2017. Exploring mixed-scale gesture interaction. In *ACM SIGGRAPH Asia 2017 Posters*. 1–2.

[10] Luigi Gallo, Alessio Pierluigi Placitelli, and Mario Ciampi. 2011. Controller-free exploration of medical image data: Experiencing the Kinect. In *Proc. CBMS*. 1–6.

[11] Brook Galna, Gillian Barry, Dan Jackson, Dadirayi Mhiripiri, Patrick Olivier, and Lynn Rochester. 2014. Accuracy of the Microsoft Kinect sensor for measuring movement in people with Parkinson's disease. *Gait & posture* 39, 4 (2014), 1062–1068.

[12] Nils Y. Hammerla, Shane Halloran, and Thomas Plotz. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables. In *Proc. IJCAI*. 1533–1540.

[13] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2010. 3D Convolutional Neural Networks for Human Action Recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35, 495–502.

[14] Bryce Kellogg, Vamsi Talla, and Shyamnath Gollakota. 2014. Bringing gesture recognition to all devices. In *Proc. NSDI*. 303–316.

[15] Jungsoo Kim, Jiasheng He, Kent Lyons, and Thad Starner. 2007. The gesture watch: A wireless contact-free gesture based wrist interface. In *Proc. ISWC*. 15–22.

[16] Tarmo Koppel, Andrei Shishkin, Heldur Haldre, N. Toropov, and Piia Tint. 2017. Reflection and Transmission Properties of Common Construction Materials at 2.4 GHz Frequency. *Energy Procedia* 113 (2017), 158–165.

[17] Sven Kratz and Michael Rohs. 2009. HoverFlow: expanding the design space of around-device interaction. In *Proc. MobileHCI*. 1–8.

[18] Christine Kühnel, Tilo Westermann, Fabian Hemmert, Sven Kratz, Alexander Müller, and Sebastian Möller. 2011. I'm home: Defining and evaluating a gesture set for smart-home control. *Int. J. Hum. Comput. Stud.* 69, 11 (2011), 693–704.

[19] David R Large, Kyle Harrington, Gary Burnett, and Orestis Georgiou. 2019. Feel the noise: Mid-air ultrasound haptics as a novel human-vehicle interaction paradigm. *Appl. Ergon.* 81 (2019), 102909.

[20] Luis A. Leiva, Matjaž Kljun, Christian Sandor, and Klen Čopič Pucihar. 2020. The Wearable Radar: Sensing Gestures Through Fabrics. In *Proc. MobileHCI*. 1–4.

[21] Yi Li. 2012. Hand gesture recognition using Kinect. In *Proc. CSAE*. 196–199.

[22] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Trans. Graphics* 35, 4 (2016), 1–19.

[23] Gan Lu, Lik-Kwan Shark, Geoff Hall, and Ulrike Zeshan. 2012. Immersive manipulation of virtual objects through glove-based hand gesture interaction. *Virtual Reality* 16, 3 (2012), 243–252.

[24] Daniel Maturana and Sebastian Scherer. 2015. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. 922–928.

[25] Jess McIntosh, Mike Fraser, Paul Worgan, and Asier Marzo. 2017. DeskWave: Desktop Interactions Using Low-Cost Microwave Doppler Arrays. In *Proc. CHI EA*. 1885–1892.

[26] Andre Mewes, Bennet Hensen, Frank Wacker, and Christian Hansen. 2017. Touchless interaction with software in interventional radiology and surgery: a systematic literature review. *Int. J. Comput. Assist. Radiol. Surg.* 12, 2 (2017), 291–305.

[27] Pranav Mistry and Pattie Maes. 2009. SixthSense: a wearable gestural interface. In *ACM SIGGRAPH ASIA 2009 Art Gallery & Emerging Technologies: Adaptation*. 85–85.

[28] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Kari Pulli. 2015. Short-range FMCW monopulse radar for hand-gesture sensing. In *Proc. RadarCon*. 1491–1496.

[29] Ismail Nasr, Reinhard Jungmaier, Ashutosh Baheti, Dennis Noppeney, Jagjit S Bal, Maciej Wojnowski, Emre Karagozler, Hakim Raja, Jaime Lien, Ivan Poupyrev, et al. 2016. A highly integrated 60 GHz 6-channel transceiver with antenna in package for smart sensing and short-range communications. *IEEE J. Solid-State Circuits* 51, 9 (2016), 2066–2076.

[30] J.Y.H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *Proc. CVPR*.

[31] Kenton O'Hara, Gerardo Gonzalez, Abigail Sellen, Graeme Penney, Andreas Varnavas, Helena Mentis, Antonio Criminisi, Robert Corish, Mark Rouncefield, Neville Dastur, et al. 2014. Touchless interaction in surgery. *Commun. ACM* 57, 1 (2014), 70–77.

[32] Francisco J. Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16, 1 (2016).

[33] Michael Otero. 2005. Application of a continuous wave radar for human gait recognition. In *Proc. SPIE*, Vol. 5809. 538–548.

[34] C.H. Oxley, J. Williams, R. Hopper, H. Flora, D. Eibeck, and C. Alabaster. 2007. Measurement of the reflection and transmission properties of conducting fabrics at milli-metric wave frequencies. *IET Science, Measurement & Technology* 1, 3 (2007), 166–169.

[35] Sameera Palipana, Dariush Salami, Luis A. Leiva, and Stephan Sigg. 2021. Pantomime: Mid-Air Gesture Recognition with Sparse Millimeter-Wave Radar Point Clouds. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1 (2021), 27:1–27:27.

[36] Joseph Paradiso, Craig Abler, Kai-yuh Hsiao, and Matthew Reynolds. 1997. The magic carpet: physical sensing for immersive environments. In *Proc. CHI EA*. 277–278.

[37] Joseph A Paradiso. 1999. The brain opera technology: New instruments and gestural sensors for musical interaction and performance. *J. New Music Res.* 28, 2 (1999), 130–149.

[38] François Petitjean, Alain Ketterlin, and Pierre Gançarski. 2011. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognit.* 44, 3 (2011), 678–693.

[39] Carl A Pickering, Keith J Burnham, and Michael J Richardson. 2007. A research study of hand gesture recognition technologies and applications for human vehicle interaction. In *Proc. IET Automotive Electronics*. 1–15.

[40] M PourMousavi, M Wojnowski, R Agethen, R Weigel, and A Hagelauer. 2013. Antenna array in eWLB for 61 GHz FMCW radar. In *Proc. APMC*. 310–312.

[41] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. 2013. Whole-home gesture recognition using wireless signals. In *Proc. MobiCom*. 27–38.

[42] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. 2017. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proc. CVPR*. 652–660.

[43] Tauhidur Rahman, Alexander T Adams, Ruth Vinisha Ravichandran, Mi Zhang, Shwetak N Patel, Julie A Kientz, and Tanzeem Choudhury. 2015. Dopplesleep: A contactless unobtrusive sleep sensing system using short-range doppler radar. In *Proc. Ubicomp*. 39–50.

[44] Siddharth S Rautaray and Anupam Agrawal. 2011. Interaction with virtual game through hand gesture recognition. In *Proc. IMPACT*. 244–247.

[45] Richard Hartless Richard Rudd , Ken Craig , Martin Ganley. 2014. *Building Materials and Propagation*. Technical Report September. 40 pages.

[46] Andreas Riener, Alois Ferscha, Florian Bachmair, Patrick Hagmüller, Alexander Lemme, Dominik Muttenthaler, David Pühringer, Harald Rogner, Adrian Tappe, and Florian Weger. 2013. Standardization of the in-car gesture interaction space. In *Proc. AutomotiveUI*. 14–21.

[47] Marco Roccetti, Gustavo Marfia, and Angelo Semeraro. 2012. Playing into the wild: A gesture-based interface for gaming in public spaces. *J. Vis. Commun. Image Represent.* 23, 3 (2012), 426–440.

[48] Christian Sandor and Hiraku Nakamura. 2018. SoliScratch: A Radar Interface for Scratch DJs. In *Proc. ISMAR-Adjunct*. 427–427.

[49] R. W. Schafer. 2011. What Is a Savitzky-Golay Filter? [Lecture Notes]. *IEEE Signal Process. Mag.* 28, 4 (2011), 111–117.

[50] Gözel Shakeri, John H Williamson, and Stephen Brewster. 2018. May the force be with you: Ultrasound haptic feedback for mid-air gesture interaction in cars. In *Proc. AutomotiveUI*. 1–10.

[51] Merrill Ivan Skolnik et al. 1980. *Introduction to radar systems*. Vol. 3. McGraw-hill New York.

[52] Jie Song, Gábor Sörös, Fabrizio Pece, Sean Ryan Fanello, Shahram Izadi, Cem Keskin, and Otmar Hilliges. 2014. In-air gestures around unmodified mobile devices. In *Proc. UIST*. 319–329.

[53] Peng Song, Wooi Boon Goh, William Hutama, Chi-Wing Fu, and Xiaopei Liu. 2012. A handle bar metaphor for virtual object manipulation with mid-air interaction. In *Proc. CHI*. 1297–1306.

[54] Thad Starner, Jake Auxier, Daniel Ashbrook, and Maribeth Gandy. 2000. The gesture pendant: A self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring. In *Proc. ISWC*. 87–94.

[55] Ivan E Sutherland. 1964. Sketchpad a man-machine graphical communication system. *Simulation* 2, 5 (1964), R–3.

[56] Ivan E Sutherland. 1965. The ultimate display. In *Proc. IFIP Congress*. 506–508.

[57] John W. Tukey. 1977. *Exploratory data analysis*. Addison-Wesley Publishing Co.

[58] Wouter Van Vlaenderen, Jens Brulmans, Jo Vermeulen, and Johannes Schöning. 2015. Watchme: A novel input method combining a smartwatch and bimanual interaction. In *Proc. CHI EA*. 2091–2095.

[59] Klen Čopič Pucihar, Christian Sandor, Matjaž Kljun, Wolfgang Huerst, Alexander Plopski, Takafumi Taketomi, Hirokazu Kato, and Luis A. Leiva. 2019. The Missing Interface: Micro-Gestures on Augmented Objects. In *Proc. CHI EA*. 1–6.

[60] Qian Wan, Yiran Li, Changzhi Li, and Ranadip Pal. 2014. Gesture recognition for smart home applications using portable radar sensors. In *Proc. EMBS*. 6414–6417.

[61] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and Otmar Hilliges. 2016. Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. In *Proc. UIST*. 851–860.

[62] Yazhou Wang and Aly E Fathy. 2011. Micro-Doppler signatures for intelligent human gait recognition using a UWB impulse radar. In *Proc. APSURSI*. 2103–2106.

[63] Teng Wei and Xinyu Zhang. 2015. mTrack: High-precision passive tracking using millimeter wave radios. In *Proc. MobiCom*. 117–129.

[64] Huiyue Wu and Jianmin Wang. 2012. User-defined body gestures for TV-based applications. In *Proc. ICDH*. 415–420.

[65] LI Yang, Jin Huang, TIAN Feng, WANG Hong-An, and DAI Guo-Zhong. 2019. Gesture interaction in virtual reality. *Virtual Reality & Intelligent Hardware* 1, 1 (2019), 84–112.

[66] Hui-Shyong Yeo, Gergely Flamich, Patrick Schrempf, David Harris-Birtill, and Aaron Quigley. 2016. RadarCat: Radar Categorization for Input & Interaction. In *Proc. UIST*. 833–841.

[67] Renyuan Zhang and Siyang Cao. 2018. Real-time human motion behavior detection via CNN using mmWave radar. *IEEE Sens. Lett.* 3, 2 (2018), 1–4.

[68] Chen Zhao, Ke-Yu Chen, Md Tanvir Islam Aumi, Shwetak Patel, and Matthew S Reynolds. 2014. SideSwipe: detecting in-air gestures around mobile devices using actual GSM signal. In *Proc. UIST*. 527–534.

[69] M. Zhao, T. Li, M. A. Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi. 2018. Through-Wall Human Pose Estimation Using Radio Signals. In *Proc. CVPR*. 7356–7365.

[70] Yan Zhuang, Chen Song, Aosen Wang, Feng Lin, Yiran Li, Changzhan Gu, Changzhi Li, and Wenyao Xu. 2015. SleepSense: Non-invasive sleep event recognition using an electromagnetic probe. In *Proc. BSN*. 1–6.

A ALTERNATIVE MODEL ARCHITECTURE

Table 4. Linear regression models (LRMs) to predict performance of 3D CNN for different gesture classifiers trained on both robot and human data with different gestures sets. The results show strong correlation across all conditions.

	Single LRM Perf ~ Trans. coef. limited (F3)							Multiple LRM Perf ~ 54 coefficients								Single LRM Perf ~ Trans. coef. limited (F3)						
Classifier (Table 1)	**Materials**	**DoF**	**RMSE**	**MAE**	**R^2**	**Adj. R^2**	**p-value**	**Materials**	**DoF**	**RMSE**	**MAE**	**R^2**	**Adj.R^2**	**p-value**	**Classifier (Table 1)**	**Materials**	**DoF**	**RMSE**	**MAE**	**R^2**	**Adj.R^2**	**p-value**
Robot A	75	73	8.69	7.02	0.78	0.78	<.001	75	20	5.72	2.28	0.97	0.91	<.001	Humans A	6	4	7.16	5.27	0.86	0.83	<.01
Robot B	75	73	10.60	8.87	0.72	0.72	<.001	75	20	7.37	2.97	0.96	0.87	<.001	Humans B	6	4	6.85	5.45	0.85	0.82	<.01
Robot C	75	73	13.00	10.20	0.59	0.58	<.001	75	20	12.00	4.95	0.90	0.64	<.01	Humans C	6	4	7.12	5.58	0.86	0.83	<.01
Robot D	75	73	11.10	8.87	0.56	0.56	<.001	75	20	7.23	2.92	0.95	0.81	<.001	Humans D	6	4	7.34	5.49	0.90	0.88	<.01
Robot E	75	73	11.60	9.61	0.71	0.70	<.001	75	20	8.93	3.69	0.95	0.82	<.001	Humans E	6	4	9.07	6.66	0.80	0.76	<.05
Robot F	75	73	9.05	6.99	0.78	0.77	<.001	75	20	7.15	2.82	0.96	0.86	<.001	Humans F	6	4	6.85	5.26	0.87	0.84	<.01
Robot G	75	73	8.19	6.65	0.81	0.81	<.001	75	20	7.61	3.14	0.96	0.84	<.001	Humans H	6	4	9.45	7.11	0.88	0.84	<.01
Mean			10.32	8.32	0.71	0.70				8.00	3.25	0.95	0.82		**Mean**			7.69	5.83	0.86	0.83	

Acronyms: *Perf: (ACC + AUC)/2 DoF: Degrees of Freedom RMSE: Room Mean Square Error MAE: Mean Absolute Error R^2: R-squared Adj. R^2: Adjusted R^2*

Previous work has shown that a spatio-temporal 3D CNN (Conv3D) architecture is an effective tool for accurate action recognition of image sequences [13, 24]. Since Soli provides a sequence of Doppler images through time, we developed a custom Conv3D architecture (Figure 9) as an alternative to our hybrid CNN+LSTM architecture (Figure 4).

This model processes each sequence of gesture images with a stack of four Conv3D blocks to extract feature maps. Each block is composed of a Conv3D layer followed by a max pooling 3D layer and a spatial dropout layer. The Conv3D layers have $32 \times 64 \times 128 \times 256$ units each with $2 \times 2 \times 2$ filters and use LeakyReLU as activation function. The max pooling layers (pool size of 2) downsample the feature maps across a 3D volume and the dropout layer (drop rate of 0.5) allows to prevent overfitting. Then, there is a fully-connected layer with 512 units followed by a softmax layer for classification.

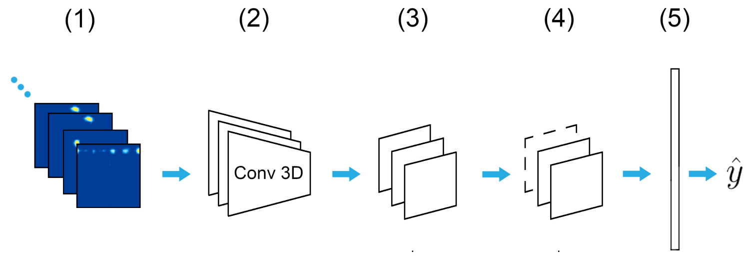


Fig. 9. Conv3D deep learning model architecture. Range Doppler images (1) are processed with a Conv3D layer (2) that extracts feature maps followed by 3D maxpooling (3) and spatial dropout (4) layers. Layers (2) to (4) are stacked in blocks of 32, 64, 128, and 256 units. Then a fully connected layer creates the feature vectors which are fed into a softmax layer (5) for class prediction ($\hat{y}$).

This model is trained with the Adam optimiser with learning rate $\eta = 0.001$ and decay rates $\beta_1 = 0.9, \beta_2 = 0.999$. As in our hybrid model training (Section 4.5), we set the maximum number of epochs to 200, and also set an early stopping criteria of 50 epochs, which means that training stopped if the validation loss did not improve after 50 consecutive epochs, and the best model weights were retained.

We built 7 classifiers for each data source (robot and humans), totaling 14 classifiers. This was done in the same way as for our hybrid CNN+LSTM architecture (see Table 1). Again, all classifiers were built for different gesture combinations (6/6 and all possible combinations of 5/6 gestures).

Table 4 shows strong correlation across all conditions, however RMSE and MAE are higher for the Conv3D architecture when compared to the hybrid architecture (see Table 2). Nevertheless, these results support our claim that our method of modeling material performance is actually agnostic to the model architecture.

Table 5 compares our two deep learning architectures considered for *all* the materials in the catalogue, using robot data. As can be seen, the hybrid CNN+LSTM model architecture achieves better performance on average. Therefore, it is preferred for measuring signal distortions with our proposed indirect method. Note that some materials induce higher signal degradation, thereby lowering recognition performance, and the Conv3D architecture is more sensitive than the hybrid CNN+LSTM architecture in this regard.

Table 5. Comparison of gesture recognition performance across all materials in our catalogue. We report mean ± standard deviation, followed by 95% confidence intervals.

Metric	**CNN+LSTM model**		**Conv3D model**	
Acc. (%)	73.62 ± 4.95	[68.67, 78.56]	66.14 ± 5.25	[60.89, 71.39]
AUC (%)	84.16 ± 2.97	[81.19, 87.13]	79.77 ± 3.15	[76.62, 82.92]