

Touch2Know: Dexterous In-Hand Liquid Sensing via Mutual-Capacitance Fingerprints

Ruixiang Deng¹, *Student Member, IEEE*, Zhegong Shangguan^{2†}, *Member, IEEE*,
Yang Hu¹, *Student Member, IEEE*, Haozheng Bai¹, *Student Member, IEEE*,
Tingcheng Li³, Angelo Cangelosi^{2†}, *Senior Member, IEEE* and Wuqiang Yang^{1,*}, *Fellow, IEEE*

Abstract—Robots and people often need to know what liquid is inside a sealed bottle before acting, but vision can fail for opaque or visually similar contents and continuous cameras may be undesirable in shared spaces. We present Touch2Know, a dexterous-hand sensing system that identifies liquid type and container type through the container wall during normal in-hand grasping, without opening the bottle. Touch2Know integrates five low-profile flex-PCB electrodes with guarded and shielded routing on a multi-finger hand, and measures multi-channel mutual capacitance to reduce common-mode drift and grounding sensitivity. An event-driven adaptive grasp controller stabilizes contact and limits over-compression across PET and glass bottles. For learning, we propose CapTF (Capacitive Temporal Fusion), a CNN-Transformer model that jointly predicts (i) liquid class and (ii) container type from 4-channel capacitive time series. We collect a dataset spanning 19 liquid classes (11 household liquids plus salt-water and sugar-water concentration series) and 4 container types over four days. The task is fine-grained, including visually ambiguous cases and liquids with closely related dielectric/conductive responses through container walls. With per-trial baseline-subtracted signals, CapTF achieves 95.07%/99.34% liquid/container accuracy under a balanced split and 90.00%/98.16% on the Day-4 holdout test, outperforming LSTM, XGBoost, and a vanilla Transformer in liquid classification. Code is available at: https://github.com/mrsgzg/Capacitive_Temporal_Fusion.

I. INTRODUCTION

Many liquids are stored and used in sealed containers, yet both people and robots often need to know what is inside before they act. This motivates a system-level capability: liquid perception that works through the container wall during normal handling, without opening the container and without strong assumptions about the environment. Vision alone is not always dependable in these cases, and continuous camera sensing can be undesirable in assistive or shared spaces. A practical solution should therefore couple perception with manipulation so that reliable liquid evidence can be gathered while the container is being held.

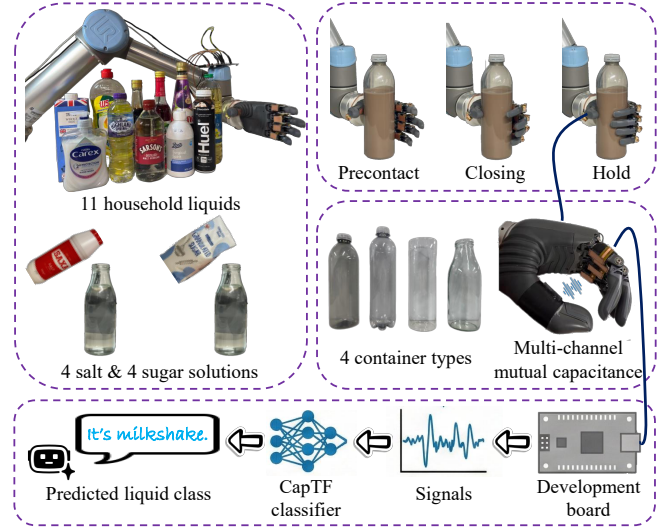


Fig. 1: Touch2Know overview. A UR5-mounted dexterous hand performs an enveloping grasp and records 4-channel mutual-capacitance signals through sealed containers using five on-hand flex-PCB electrodes. The dataset covers 19 liquid classes and four container types over four days.

A dexterous hand is well suited for this goal. Compared with two-finger grasps, a multi-finger enveloping grasp can stabilize containers with different diameters and materials and maintain distributed contact around the bottle. This contact pattern provides multiple sensing viewpoints at once and makes the measurement less sensitive to small placement errors. It also matches real handling: the robot can pick up a bottle from a table or accept one from a user, hold it securely, and sense it in-hand.

In this paper, we present Touch2Know, a dexterous-hand sensing system that identifies both liquid class and container type using mutual-capacitance signals measured through sealed containers (Fig. 1), without opening the container and regardless of whether the bottle is transparent or opaque. We integrate low-profile flex-PCB electrodes on a dexterous hand and acquire multi-channel mutual-capacitance responses with a Ruby Development Board. An adaptive grasp controller stabilizes contact conditions and limits excessive force, improving repeatability across plastic and glass containers. We focus on classification: (i) 19 liquid classes (11 household liquids plus 4 salt-water and 4 sugar-water solutions treated as discrete classes), and (ii) 4 container types.

The contributions of this paper are:

- 1) A dexterous-hand mutual-capacitance sensing system

*Corresponding author. (wuqiang.yang@manchester.ac.uk)

[†]This work was supported by the ERC eTALK Project (Grant Reference: EP/Y029534/1).

¹Ruixiang Deng, Yang Hu, Haozheng Bai, and Wuqiang Yang are with the Department of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, U.K.

²Zhegong Shangguan and Angelo Cangelosi are with the Department of Computer Science, The University of Manchester, Manchester M13 9PL, U.K.

³Tingcheng Li is with the School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China.

for in-hand liquid and container classification through sealed containers, leveraging multi-channel spatial sampling during an enveloping grasp.

- 2) A low-cost, scalable flex-PCB electrode design with shielded routing and detachable SMA-coax connections for repeatable integration and maintenance.
- 3) An event-driven adaptive grasp controller that stabilizes contact and limits over-compression, improving repeatability across plastic and glass containers.
- 4) We propose CapTF (Capacitive Temporal Fusion), a hybrid CNN-Transformer model for liquid and container classification from mutual-capacitance sequences.
- 5) A fine-grained dataset of 19 liquid classes (11 household liquids plus 4 salt-water and 4 sugar-water solutions treated as discrete classes) and 4 container types collected over four days, evaluated with balanced and Day-4 holdout splits, showing the benefit of per-trial baseline-subtracted delta features.

II. RELATED WORK

A. Liquid Detection with Robots

Robotic liquid understanding has mainly used (i) vision, (ii) interaction-based probing, and (iii) non-visual end-effector sensing. Vision methods use RGB/RGB-D or multi-view cues (e.g., surface appearance, refraction, and geometry) to estimate fill level or infer contents [1], [2]. In real settings, they often degrade under occlusion, glare, low contrast for transparent liquids, and opaque containers, and continuous cameras can be undesirable in shared spaces [3], [4], [5]. Another line of work applies actions such as shaking [6], tilting [7], or external forcing [8], and then classifies liquids using vibration/haptic responses [9] or audio signals [10]. These pipelines typically add extra action stages and can be less stable when liquids have similar dynamics. Non-visual end-effector sensing (e.g., spectroscopy [11], thermal sensing [9], millimeter-wave sensing [12], and electrical impedance tomography [13]) can reduce reliance on vision, but it often increases hardware cost and calibration burden.

B. Reliable Liquid Sensing with a Dexterous Hand

Most liquid sensing systems rely on two-finger grippers or fixture-like end effectors to simplify contact geometry and control the sensing pose [7], [11], [13], [14]. A dexterous hand can instead form an enveloping grasp that adapts to different diameters and improves stability [15], while also providing distributed contact that enables multi-point spatial sampling around the container [16]. However, relatively few studies embed liquid sensing into dexterous grasping; existing examples mainly target liquid-level or related in-hand estimates [17], [9]. Integrating sensing and stable grasping in one platform can avoid a separate measurement procedure and better match home deployment needs [18].

C. Capacitive Tactile Sensor for Liquid Detection

Capacitive tactile sensing is widely used for end-effector liquid detection because thin, flexible electrodes are easy

to integrate and can respond to electrical property changes through the container wall [7], [14], [19], [6], [17]. It has been used for fill-level estimation [20] and for content or concentration-related recognition [21], [14]. Existing electrode configurations mainly include self-capacitance [14] and mutual capacitance [21], [22], [23], [24], [25]. Self-capacitance is simple and easy to implement, but it is highly sensitive to environmental changes and grounding conditions, as it measures a single-ended electrode-to-ground coupling. This makes the signal prone to common-mode disturbances (e.g., nearby conductive objects, humidity, and robot body motion), causing drift and reduced repeatability, and ultimately limiting robustness in real-world robotic settings [26]. Mutual capacitance uses a transmit-receive electrode pair to form a more well-defined coupling path through the target, which reduces dependence on grounding and suppresses common-mode environmental disturbances [22], [27]. This enables stable multi-channel comparative features and improves discrimination for material and content sensing at robotic end effectors [21], [23].

III. METHODOLOGY

A. Capacitive Tactile Sensor Design, Fabrication, and Integration

Fig. 2 summarizes the full workflow from flex-PCB fabrication to on-hand integration and mutual-capacitance data acquisition.

a) Flex-PCB substrate and fabrication.

The electrodes were fabricated on a double-sided copper-polyimide laminate (Cu/PI/Cu) with a stack-up of 18 μm Cu / 25 μm PI / 18 μm Cu (Fig. 2a). This structure is flexible and conductive, so it can conform to curved finger surfaces. The layout was designed in *Altium Designer* and manufactured as a flex-PCB (Fig. 2b). The flex-PCB process provides low-cost, repeatable geometry and supports fast design iteration.

b) Electrode structure.

Fig. 2c shows the electrode and the layer functions. On the front side (object-facing), a rectangular copper pad is the sensing electrode. The pad is surrounded by a closed GND guard ring on the same side (width = 1 mm). This guard provides a local reference and limits the lateral fringing field. It directs coupling toward the container and improves repeatability. On the back side (routing and shielding), the sensing pad connects to a narrow trace that serves as the electrode signal feed to the connector. A grounded copper shield (GND) surrounds the trace. The shield reduces coupling into the finger body and internal wires. It also lowers parasitic capacitance and cross-talk across channels. Together, the front guard ring and the back shield make the field more stable for multi-channel mutual-capacitance sensing.

c) SMA interface with coax-like launch.

At the connector area, the PCB uses a coax-like hole pattern (Fig. 2d): the center hole is the signal, and the four outer holes are GND; an SMA connector is inserted from the back side, its pins are soldered on the front side to form the solder joints, and the joint region is then reinforced with

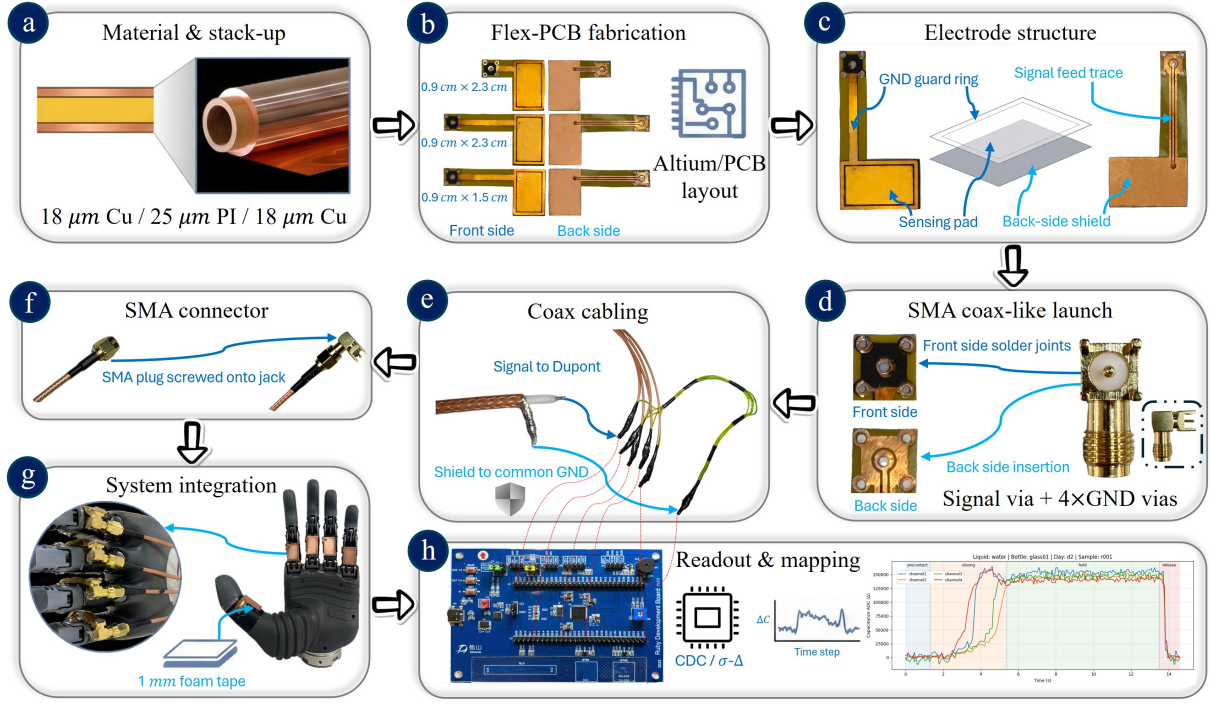


Fig. 2: Flex-PCB electrode workflow and system integration for multi-channel mutual-capacitance sensing.

hot-melt glue for mechanical stability and strain relief. This geometry provides a short return path and reduces connector parasitics, supporting stable mutual-capacitance readout.

d) Coax cabling and SMA connector.

As shown in Fig. 2e-f, each channel uses a short RF coax run between the flex-PCB SMA jack and an SMA plug. The cable inner conductor carries the electrode signal to the readout wiring, while the cable shield is tied to a shared ground to maintain a continuous shielding path and reduce relative drift. The detachable SMA pair simplifies replacement.

e) System integration.

Two electrode sizes were used to match finger geometry: (i) thumb + middle + ring: $0.9\text{ cm} \times 2.3\text{ cm}$; (ii) index + pinky: $0.9\text{ cm} \times 1.5\text{ cm}$. The flex-PCB is mounted at the second joint, then wraps to the dorsal side and routes toward the back of the hand. The flex tail and cable are fixed using hot-melt glue for strain relief. (Fig. 2g). This location provides stable contact in an enveloping grasp and avoids interference with the fingertip 3-axis force sensor. Each electrode was bonded using 1 mm sponge double-sided tape behind the back-side shield. The foam adds compliance and improves contact. It also increases distance to internal conductors, which can reduce parasitic capacitance and improve baseline stability.

f) Signal readout & mapping.

The five electrodes are connected to the Ruby board as: thumb $\rightarrow$ G35, index $\rightarrow$ G07, middle $\rightarrow$ G12, ring $\rightarrow$ G25, pinky $\rightarrow$ G20.

Mutual-capacitance signals are recorded using a development board from Beijing Tashan Technology Co., Ltd.

equipped with a 24-bit $\sigma\text{-}\delta$ capacitive-to-digital converter (CDC) (Fig. 2h). In this work, the board streams synchronized multi-channel capacitance samples to the host at a fixed rate, producing time series. Fig. 2h shows a representative dataset trial (water in glass01, Day 2) plotted as baseline-referenced delta capacitance ADC (raw minus the per-trial environment baseline mean). The four channels increase during closing and settle to a stable delta plateau during hold before returning toward zero on release, illustrating the repeatable in-hand fingerprint after baseline calibration.

B. Mutual-Capacitance Sensing Principle

We measure *mutual capacitance* between a transmit (Tx) electrode on the thumb and four receive (Rx) electrodes on the other fingers during a stable enveloping grasp. With electrodes conforming to the bottle wall, the dominant electric-field path traverses the container wall and couples into the liquid. Liquids with different relative permittivity and conductivity therefore produce distinct multi-channel temporal signatures.

In a simplified Tx-Rx model, the induced charge at Rx is

$$Q_{\text{rx}}(t) = C_m V_{\text{tx}}(t), \quad (1)$$

where C_m is the mutual capacitance and $V_{\text{tx}}(t) = V_0 \sin(\omega t)$ is the sinusoidal excitation (amplitude V_0 , angular frequency ω) applied to Tx. The corresponding Rx current is

$$i_{\text{rx}}(t) = \frac{dQ_{\text{rx}}}{dt} = \omega C_m V_0 \cos(\omega t), \quad (2)$$

so C_m can be estimated from the measured peak current I_{pk} as

$$C_m \approx \frac{I_{\text{pk}}}{\omega V_0}. \quad (3)$$

To connect C_m with the bottle-liquid structure, we approximate the coupling as an effective layered capacitance (wall + liquid),

$$C_m \approx \frac{\varepsilon_0 A_{\text{eff}}}{\frac{d_w}{\varepsilon_w} + \frac{d_\ell}{\varepsilon_{\ell, \text{eff}}}}, \quad (4)$$

where ε_0 is vacuum permittivity, A_{eff} is an effective coupling area, d_w and ε_w denote wall thickness and relative permittivity, and d_ℓ and $\varepsilon_{\ell, \text{eff}}$ denote an effective liquid path length and permittivity under fringing fields. For lossy liquids, conductivity σ contributes via the complex permittivity

$$\varepsilon_\ell^*(\omega) = \varepsilon_\ell'(\omega) - j \frac{\sigma}{\omega \varepsilon_0}. \quad (5)$$

Using the thumb as Tx and the remaining fingers as Rx yields a 4-channel sequence $\mathbf{C}_m(t)$ for learning-based classification.

C. Adaptive Grasping Algorithm

We use a dexterous hand from BrainCo with 6 actuated joints and 10 total DoF, communicating with the host PC via RS485. The thumb has two actuated DoF (abduction/adduction and flexion/extension), while the index, middle, ring, and pinky each have one actuated flexion/extension DoF plus one passive flexion DoF, enabling underactuated wrap-around grasps. The hand integrates capacitive 3D fingertip force sensors from Beijing Tashan Technology Co., Ltd. that estimate normal and tangential forces (0–25 N range, ~ 0.1 N resolution, 5% FS accuracy). The thumb provides two sensing channels, whereas the other fingers provide three channels each. Each channel uses a three-layer stack (contact layer, silicone-rubber dielectric, and electrode layer); external forces deform the dielectric and change the measured capacitance, which is converted by the sensor electronics into normal and tangential force estimates.

We use an event-driven controller to obtain a repeatable enveloping grasp so that the second-joint capacitive electrodes achieve stable contact while avoiding over-compression. The controller follows a state machine: pre-contact $\rightarrow$ closing $\rightarrow$ hold $\rightarrow$ release $\rightarrow$ done; the hold state freezes the posture for a fixed sensing window. We keep `thumb_aux` at its maximum to improve opposition geometry and align the thumb with the index across bottle diameters.

In closing, each finger j (thumb, index, middle, ring, and pinky) maintains a mode in `{active, holding}` and uses a robust normal-force feature

$$F_{n, \max}^{(j)}(k) = \max_i F_{n, i}^{(j)}(k). \quad (6)$$

If `active`, the finger advances by a small step

$$p_j(k+1) = p_j(k) + \Delta p, \quad p_j \in [0, 100]. \quad (7)$$

Contact is detected with hysteresis: when $F_{n, \max}^{(j)}(k) > F_{\text{contact}}$, finger j switches to `holding`; when `holding` and later $F_{n, \max}^{(j)}(k) < F_{\text{release}}$ (with $F_{\text{release}} < F_{\text{contact}}$), it may reactivate. After the first contact event we bound additional closure to limit compression:

$$p_j \leq p_{j, \text{first}} + \Delta p_{\text{extra}}. \quad (8)$$

We also stop a finger if it stalls (command increases but readback does not), and only transmit new position commands when the target posture changes. Closing ends when all fingers are non-active, after which the controller records during `hold` and executes a one-shot `release`.

D. Capacitive Temporal Fusion Method

We formulate in-hand liquid sensing as a *multi-task* time-series classification problem. A single grasp trial produces a C -channel mutual-capacitance sequence from which the model must simultaneously predict (i) the liquid class (19 classes: 11 household liquids plus 4 salt-water and 4 sugar-water concentration levels) and (ii) the container type (4 classes). Solving both tasks jointly is motivated by the observation that the capacitive response depends on both the liquid and the container wall, and a shared representation can capture their interaction more efficiently than two independent models. Because the signal evolves through distinct grasp phases and contains both local transient patterns and long-range temporal dependencies, we propose **CapTF** (**Capacitive Temporal Fusion**), a multi-task CNN-Transformer architecture that shares a convolutional-attentional backbone and branches into task-specific classification heads.

1) Problem Formulation

Let a single grasp trial produce a multi-channel time series $\mathbf{X} \in \mathbb{R}^{C \times L}$, where $C = 4$ is the number of Rx mutual-capacitance channels and L is the sequence length. Each trial is associated with two ground-truth labels: a liquid label $y^{\text{liq}} \in \{1, \dots, K_{\text{liq}}\}$ and a container label $y^{\text{bot}} \in \{1, \dots, K_{\text{bot}}\}$. The goal is to learn a mapping f_θ that produces two logit vectors from the same input:

$$\mathbf{o}^{\text{liq}}, \mathbf{o}^{\text{bot}} = f_\theta(\mathbf{X}), \quad (9)$$

where $\mathbf{o}^{\text{liq}} \in \mathbb{R}^{K_{\text{liq}}}$ and $\mathbf{o}^{\text{bot}} \in \mathbb{R}^{K_{\text{bot}}}$. The predicted class probabilities are obtained via softmax:

$$p(y^\tau = k | \mathbf{X}) = \frac{\exp(o_k^\tau)}{\sum_{j=1}^{K_\tau} \exp(o_j^\tau)}, \quad \tau \in \{\text{liq}, \text{bot}\}, \quad (10)$$

and the predicted labels are $\hat{y}^\tau = \arg \max_k p(y^\tau = k | \mathbf{X})$.

The model is trained with a weighted multi-task cross-entropy loss:

$$\mathcal{L} = \lambda_{\text{liq}} \mathcal{L}_{\text{CE}}^{\text{liq}} + \lambda_{\text{bot}} \mathcal{L}_{\text{CE}}^{\text{bot}}, \quad (11)$$

where λ_{liq} and λ_{bot} control the relative importance of the two tasks. Joint training encourages the shared backbone to learn features informative for both liquid content and container geometry.

2) CapTF Architecture

The overall architecture of CapTF, illustrated in Fig. 3, consists of three stages: (i) a convolutional embedding front-end, (ii) a Transformer encoder with learnable positional encoding, and (iii) a multi-head pooling layer followed by two task-specific classification heads.

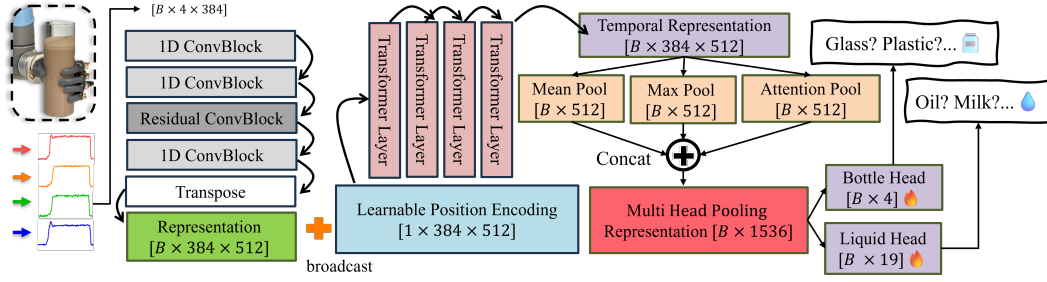


Fig. 3: CapTF architecture. A multi-scale 1-D CNN embeds the 4-channel mutual-capacitance input, a Transformer encoder with learnable positional encoding captures temporal dependencies, and multi-head pooling (mean, max, attention) aggregates the sequence into a shared representation fed to two task-specific heads for liquid and container classification.

a) Convolutional embedding.

The input tensor $\mathbf{X} \in \mathbb{R}^{B \times C \times L}$ is passed through three successive 1-D convolutional blocks that progressively expand the channel dimension ($C \rightarrow 64 \rightarrow 128 \rightarrow d$) using kernel sizes of 7, 5, and 3, respectively. Each block applies convolution, batch normalization, GELU activation, and dropout. Between the second and third blocks, a residual convolutional module with a skip connection is inserted:

$$\mathbf{H}^{(\text{res})} = \text{Conv}_2(\text{GELU}(\text{BN}(\text{Conv}_1(\mathbf{H})))) + \mathbf{H}, \quad (12)$$

which enriches local feature extraction while stabilizing gradient flow. The decreasing kernel sizes ($7 \rightarrow 5 \rightarrow 3$) allow the front-end to capture multi-scale temporal patterns: early layers extract broad trends in the capacitive response, while later layers resolve finer transient details. The output feature map $\mathbf{H} \in \mathbb{R}^{B \times d \times L}$ is transposed to $\mathbb{R}^{B \times L \times d}$ for the subsequent Transformer.

b) Learnable positional encoding.

Since self-attention is permutation-equivariant, explicit position information must be injected into the sequence. We use a learnable positional encoding $\mathbf{P} \in \mathbb{R}^{1 \times L \times d}$, initialized from $\mathcal{N}(0, 1/\sqrt{d})$ and added element-wise to the embedded sequence:

$$\tilde{\mathbf{H}} = \mathbf{H} + \mathbf{P}. \quad (13)$$

Compared with fixed sinusoidal encodings, a learnable parameterization allows the model to adapt its positional representation to the non-uniform temporal structure of the grasp phases (short precontact, variable-length closing, long hold), which are specific to our acquisition protocol and difficult to capture with a predefined frequency basis.

c) Transformer encoder.

The encoded sequence is processed by a stack of N_{enc} Transformer encoder layers with pre-layer normalization[28]. Each layer applies multi-head self-attention followed by a position-wise feed-forward network:

$$\mathbf{Z}' = \mathbf{Z} + \text{MHSA}(\text{LN}(\mathbf{Z})), \quad (14)$$

$$\mathbf{Z}'' = \mathbf{Z}' + \text{FFN}(\text{LN}(\mathbf{Z}')), \quad (15)$$

where MHSA denotes multi-head self-attention with h heads and FFN is a two-layer MLP with hidden dimension d_{ff} and GELU activation. Self-attention enables every time step to attend to the full sequence, capturing long-range dependencies

such as the transition from closing dynamics to steady-state hold signals. A final layer normalization is applied after the last encoder layer.

d) Multi-head pooling.

To aggregate the Transformer output $\mathbf{Z} \in \mathbb{R}^{B \times L \times d}$ into a fixed-size representation, we combine three complementary pooling strategies in parallel. Global mean pooling captures the average signal level across the entire grasp. Global max pooling highlights the most salient time steps, which often correspond to contact transients or peak capacitive coupling. Learned attention pooling applies a lightweight query network g_ϕ that assigns a scalar importance weight to each time step:

$$\alpha_t = \frac{\exp(g_\phi(\mathbf{Z}_t))}{\sum_{t'=1}^L \exp(g_\phi(\mathbf{Z}_{t'}))}, \quad (16)$$

where g_ϕ is a two-layer MLP ($d \rightarrow d/4 \rightarrow 1$) with GELU activation. The three pooled vectors are concatenated to form a multi-view representation:

$$\mathbf{z} = \left[\underbrace{\frac{1}{L} \sum_t \mathbf{Z}_t}_{\text{mean}}; \underbrace{\max_t \mathbf{Z}_t}_{\text{max}}; \underbrace{\sum_t \alpha_t \mathbf{Z}_t}_{\text{attention}} \right] \in \mathbb{R}^{3d}. \quad (17)$$

This multi-head design provides a richer shared representation than any single pooling method: mean pooling smooths out noise, max pooling preserves discriminative peaks, and attention pooling learns task-relevant temporal weighting.

e) Task-specific classification heads.

The shared representation $\mathbf{z}$ is fed into two independent classification heads. Each head consists of its own layer normalization, dropout, and a linear projection:

$$\mathbf{o}^{\text{liq}} = \mathbf{W}_{\text{liq}} \text{Dropout}(\text{LN}_{\text{liq}}(\mathbf{z})) + \mathbf{b}_{\text{liq}}, \quad (18)$$

$$\mathbf{o}^{\text{bot}} = \mathbf{W}_{\text{bot}} \text{Dropout}(\text{LN}_{\text{bot}}(\mathbf{z})) + \mathbf{b}_{\text{bot}}, \quad (19)$$

where $\mathbf{o}^{\text{liq}} \in \mathbb{R}^{K_{\text{liq}}}$ and $\mathbf{o}^{\text{bot}} \in \mathbb{R}^{K_{\text{bot}}}$. Using separate normalization and dropout for each head allows the two tasks to maintain different feature scales and regularization levels while still benefiting from the shared backbone. The liquid head focuses on permittivity- and conductivity-driven signal differences, while the container head attends to wall-thickness and diameter-dependent coupling variations. Both heads are trained jointly through the multi-task loss $\mathcal{L}$.

IV. EXPERIMENT

A. Datasets and Data Collection Procedure

We collected one dataset for joint liquid and container classification, covering 19 liquid classes, including 11 household liquids (water, denatured ethyl alcohol 96–99.99%, grape juice, milk, milkshake, oil, soy sauce, syrup, vinegar, handwash, and dishwashing liquid) and prepared concentration variants of salt water (2%, 4%, 6%, 8%) and sugar water (5%, 10%, 15%, 20%). This selection includes visually ambiguous cases such as adjacent concentration levels and liquids with closely related electrical properties, making the classification problem fine-grained.

All liquids were measured using four container types to capture container-dependent coupling: pet01 (hard PET, small diameter), pet02 (soft PET, large diameter), glass01 (thin glass, small diameter), and glass02 (thick glass, large diameter). All bottles were filled to 450 mL (near-full), so that the liquid level was above the highest capacitive electrode during grasping. Data were collected over four days under normal laboratory room-temperature conditions, with a mean ambient temperature of 22°C. Each day, each liquid (or concentration level) was recorded for 10 trials in each container. In total, we collected $19 \times 4 \times 10 \times 4 = 3040$ trials. This design exposes the sensing system to variations in wall material/thickness, diameter, and day-to-day setup conditions, while providing repeated measurements for reliable evaluation.

a) Per-trial acquisition workflow and logging.

Each trial follows the same sequence: initialization $\rightarrow$ environment baseline $\rightarrow$ precontact $\rightarrow$ closing $\rightarrow$ hold $\rightarrow$ release/done. At the start of each trial, the software starts a continuous logger that writes one row to the CSV at a fixed host-side rate of 25 Hz.

Before contact phases, we record a 5 s environment baseline and compute a trial-specific baseline mean from this window. Recomputing the baseline for every trial reduces sensitivity to slow drift and setup changes (e.g., cable routing, nearby objects, humidity), improving cross-trial comparability.

After baseline, the bottle is placed into the hand at a random pose without external fixtures. This avoids requiring precise placement and increases realism. The adaptive grasp controller then drives the hand toward a stable enveloping grasp. After pressing “start,” the trial records 1 s of pre-contact, then a variable-duration closing phase until stable contact is achieved, followed by an 8 s hold window, and finally release/done.

b) Sample definition, naming, and leakage prevention.

One CSV file corresponds to one complete trial, and one file is treated as one sample. Each row includes the current phase label, hand motor states, fingertip tactile readings, the latest MCU mutual-capacitance values, and the baseline-referenced delta signals computed using the trial-specific environment baseline. Filenames follow: `liq-<label>__bid-<container>__d<day>`

TABLE I: Performance comparison and delta/raw feature ablation under Balanced and Date splits. Date denotes the Day-4 holdout split. Best results are in **bold**. Bot. Acc./F1 are reported for CapTF only.

Model	Metric	Delta		Raw	
		Balanced	Date	Balanced	Date
CapTF (Ours)	Liq. Acc.	95.07%	90.00%	80.15%	68.82%
	Liq. F1	95.07%	89.89%	80.17%	68.28%
	Bot. Acc.	99.34%	98.16%	96.82%	97.24%
	Bot. F1	99.34%	98.16%	96.86%	97.26%
CapTF-Liq	Liq. Acc.	92.11%	86.51%	76.75%	65.79%
	Liq. F1	90.90%	86.52%	76.72%	65.70%
Transformer	Liq. Acc.	47.15%	45.00%	51.86%	43.29%
	Liq. F1	47.16%	45.66%	51.74%	45.66%
LSTM	Liq. Acc.	91.56%	83.68%	78.73%	67.11%
	Liq. F1	91.53%	83.72%	78.80%	66.57%
XGBoost	Liq. Acc.	79.39%	77.37%	57.35%	51.84%
	Liq. F1	79.25%	77.50%	57.31%	51.77%

`__r-<trial>.csv, <container>` is one of {pet01, pet02, glass01, glass02}, `<day>` is {d1–d4}, and `<trial>` indexes repetitions.

B. Training Details

All neural models are implemented in PyTorch 2.5.1 and trained on an NVIDIA L40 GPU with a 12-core CPU. Data splits are at the trial-file level to prevent temporal leakage. The input sequence length is $L = 384$ for all temporal models.

For CapTF, we use the AdamW optimizer with cosine-annealing learning rate schedule, batch size 64, learning rate 5×10^{-4} , weight decay 10^{-2} , gradient clipping at 1.0, and dropout 0.05. The default architecture hyperparameters are: $d = 512$, $h = 8$, $N_{\text{enc}} = 4$, and $d_{\text{ff}} = 1024$.

The vanilla Transformer baseline shares the same optimizer, learning rate, and architecture dimensions ($d = 512$, $h = 8$, 4 layers, $d_{\text{ff}} = 1024$) but omits the CNN front-end, residual blocks, and multi-head pooling. The LSTM baseline uses a 4-layer network with 512 hidden units and the same optimizer settings. XGBoost uses max depth 6, learning rate 0.1, subsample ratio 0.8, column sample ratio 0.8, and early stopping with patience 50 over 500 boosting rounds. All experiments use a fixed seed of 256 for reproducibility.

V. RESULTS

We evaluate CapTF against three baselines (vanilla Transformer, LSTM, XGBoost) under two split strategies: a *balanced split* (stratified 70/30 by liquid-container condition) and a *date split* (Days 1–3 train, Day 4 test). Both delta (baseline-subtracted) and raw features are tested. All splits are at the trial-file level to prevent leakage. We report overall accuracy and macro-averaged F1-score for liquid (19-class) and container (4-class) classification. The liquid set is fine-grained, including adjacent concentration levels and visually similar household liquids with closely related through-wall capacitive responses.

A. Liquid and Container Classification

Table I summarizes the results. Under the balanced split with delta features, CapTF achieves 95.07% liquid accu-

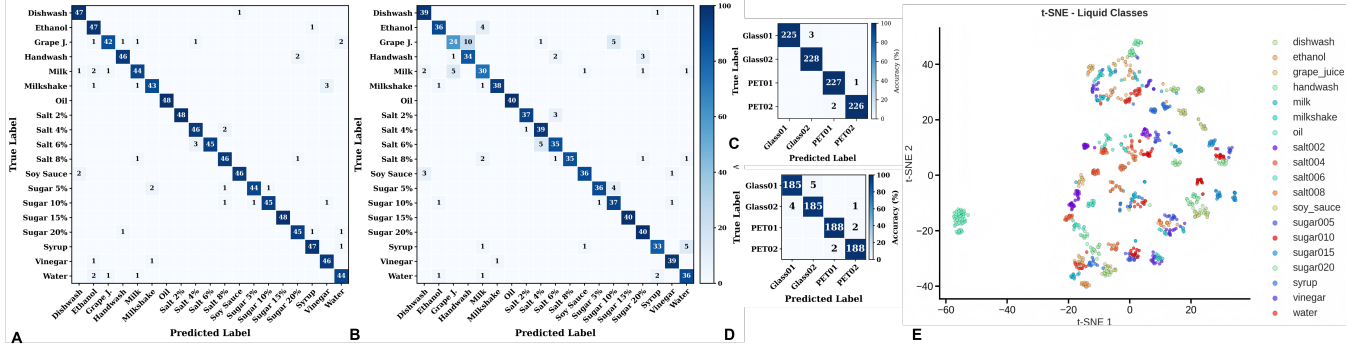


Fig. 4: Confusion matrices and representation analysis for CapTF with delta features under balanced and cross-day evaluation: (A) liquid confusion matrix (balanced split), (B) liquid confusion matrix (date split: Day-4 holdout), (C) container confusion matrix (balanced split), (D) container confusion matrix (date split: Day-4 holdout), and (E) t-SNE visualization of pooled embeddings for liquid classes.

racy and 95.07% macro-F1, outperforming LSTM (91.56%), XGBoost (79.39%), and the vanilla Transformer (47.15%). The balanced liquid confusion matrix (Fig. 4A) is strongly diagonal, indicating that most classes produce stable and separable mutual-capacitance fingerprints once per-trial drift is removed. The remaining errors cluster among water-based mixtures, visually similar household liquids, and adjacent concentration levels in the salt/sugar series, while liquids with distinctive electrical responses (e.g., oil) are nearly perfectly recognized.

The vanilla Transformer performs poorly (~ 43 – 52% across settings), suggesting that self-attention alone is insufficient for 4-channel capacitive sequences: the signal contains short contact transients and longer steady-state plateaus across grasp phases, requiring CapTF’s CNN front-end for local temporal structure and Transformer encoder for long-range phase-to-phase context.

CapTF simultaneously achieves near-ceiling container performance: 99.34% container accuracy (balanced) and 98.16% (date), as shown in Table I. The container confusion matrices (Figs. 4C–D) show only rare errors, mainly between same-material pairs (glass or PET bottles differing in thickness and compliance), with glass-type confusions increasing under the date split due to day-to-day variation.

Multi-task learning provides consistent gains for liquid classification. Compared with the single-task variant CapTF-Liq, CapTF improves liquid accuracy by 2.96 pp on the balanced split (95.07% vs. 92.11%) and 3.49 pp on the date split (90.00% vs. 86.51%). This is consistent with through-wall capacitive sensing physics: the measured coupling depends jointly on liquid properties and container wall properties. Predicting container type encourages the shared backbone to learn features that are less entangled with container-dependent offsets, improving liquid separability.

B. Delta Features vs. Raw Features

The raw-vs-delta comparison in Table I serves as a feature ablation of the per-trial baseline-subtraction step. Delta features substantially improve CapTF, LSTM, and XGBoost across both split strategies, while the vanilla Transformer shows mixed results. For CapTF, liquid accuracy increases

from 80.15% to 95.07% on the balanced split and from 68.82% to 90.00% on the date split. LSTM and XGBoost show similar trends, improving from 78.73% to 91.56% and from 57.35% to 79.39% on the balanced split, respectively, with even larger gains on the date split. These results indicate that raw mutual-capacitance measurements contain substantial trial- and day-dependent offsets driven by environmental coupling (e.g., grounding, cable routing, nearby conductors, humidity). Such offsets can dominate class-dependent variation, especially for fine-grained concentration levels and liquids with similar permittivity/conductivity when sensed through a container wall. Per-trial baseline subtraction acts as a lightweight calibration step that increases effective inter-class margins and stabilizes learning.

Container classification also benefits from delta features (99.34% vs. 96.82% on the balanced split; 98.16% vs. 97.24% on the date split), suggesting that baseline removal sharpens wall-dependent coupling differences as well, particularly for thickness/compliance variation within the same material family.

C. Generalization Across Days

The date split evaluates robustness to day-to-day changes in ambient conditions and setup (e.g., hand placement, contact pressure distribution, cable geometry, and background coupling). With delta features, CapTF drops from 95.07% (balanced) to 90.00% (date), while LSTM drops from 91.56% to 83.68% and XGBoost drops from 79.39% to 77.37%. CapTF maintains the best absolute Day-4 performance, suggesting that its multi-scale temporal embedding and global context modeling improve robustness to day-specific artifacts.

The Day-4 liquid confusion matrix (Fig. 4B) shows that most classes remain well-recognized, but errors increase for a small subset of challenging liquids and for adjacent concentration levels. Confusions are most pronounced among water-based liquids with additives (e.g., grape juice, handwash, and some sugar/salt solutions), which likely have smaller electrical separation through the container wall and are more sensitive to contact variation. Liquids with more distinctive responses (e.g., oil) remain stable on Day 4. Overall, the

results indicate that delta calibration and multi-task learning are key for robust through-wall capacitive recognition, while domain shift remains the main limiting factor for the hardest, small-margin liquid pairs.

VI. CONCLUSION

We presented Touch2Know, a dexterous in-hand sensing system for through-container liquid and container classification using multi-channel mutual-capacitance fingerprints. The system combines guarded/shielded flex-PCB electrodes, an adaptive grasp controller, and CapTF, a multi-task CNN–Transformer for 4-channel capacitive time series. On a dataset of 19 liquid classes and four container types collected over four days, CapTF achieves 95.07% liquid accuracy and 99.34% container accuracy under a balanced split, and 90.00% liquid accuracy and 98.16% container accuracy on the Day-4 holdout test. These results show that baseline-subtracted mutual-capacitance signals provide discriminative non-visual cues for in-hand liquid recognition.

Current limitations include the restricted set of bottle types and the near-full fill condition. Future work will study broader container geometries, fill-level variation, calibration transfer, open-set recognition, and integration with downstream manipulation tasks such as pouring and handover.

ACKNOWLEDGMENT

The authors gratefully thank Beijing Tashan Technology Co., Ltd. for technical support and the hardware platform, as well as CF3 and Research IT at The University of Manchester for their assistance and access to the Research VM Platform.

REFERENCES

- [1] C. Do, T. Schubert, and W. Burgard, “A probabilistic approach to liquid level detection in cups using an rgb-d camera,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2075–2080, 2016.
- [2] W. Lai, T. Zhang, T. L. Lam, and Y. Gao, “Vision-language model-based physical reasoning for robot liquid perception,” in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9652–9659, 2024.
- [3] L. Adriaens, T. Lips, M. De Coster, A. Verleysen, and F. Wyffels, “Spill: Size, pose, and internal liquid level estimation of transparent glassware for robotic bartending,” *IEEE Robotics and Automation Letters*, vol. 10, no. 12, pp. 13288–13295, 2025.
- [4] W. Zhao, F. Han, X. Qiu, X. Peng, Y. Zhao, and J. Zhang, “Research on the identification and distribution of biofouling using underwater cleaning robot based on deep learning,” *Ocean Engineering*, vol. 273, p. 113909, 2023.
- [5] S. Xu, W. Yuan, and Y. Cai, “Deep learning for velocity and range estimation of drones in urban occlusion environments,” *IEEE Transactions on Network Science and Engineering*, vol. 13, pp. 6453–6469, 2026.
- [6] H. Saal, J. Ting, and S. Vijayakumar, “Active sequential learning with tactile feedback,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Y. W. Teh and M. Titterton, eds.), vol. 9 of *Proceedings of Machine Learning Research*, (Chia Laguna Resort, Sardinia, Italy), pp. 677–684, PMLR, 13–15 May 2010.
- [7] C. Matl, R. Matthew, and R. Bajcsy, “Haptic perception of liquids enclosed in containers,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7142–7149, 2019.
- [8] S. Chitta, M. Piccoli, and J. Sturm, “Tactile object class and internal state recognition for mobile manipulation,” in *2010 IEEE International Conference on Robotics and Automation*, pp. 2342–2348, 2010.
- [9] H. Park and Y. Cha, “Liquid classification in robotic fingers with multimodal tactile sensor system,” *npj Robotics*, vol. 3, p. 33, 2025.
- [10] S. Donaher, A. Xompero, and A. Cavallaro, “Audio classification of the content of food containers and drinking glasses,” in *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 591–595, 2021.
- [11] N. Hanson, W. Lewis, K. Puthuvelil, D. Furline, A. Padmanabha, T. Padi, and Z. Erickson, “Slurp! spectroscopy of liquids using robot pre-touch sensing,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3786–3792, 2023.
- [12] D. Cao, Y. Lin, G. Ren, Y. Gao, and W. Dong, “Mmliquid: Liquid identification using mmwave,” in *Wireless Sensor Networks* (H. Ma, X. Wang, L. Cheng, L. Cui, L. Liu, and A. Zeng, eds.), (Singapore), pp. 1–18, Springer Nature Singapore, 2022.
- [13] J. Zhu, Y. J. Lee, Y. Luo, T. Xu, C. Liu, D. Rus, S. Mueller, and W. Matusik, “Liquids identification and manipulation via digitally fabricated impedance sensors,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 18164–18171, 2024.
- [14] Y. Hu, A. Gillespie, A. Padmanabha, K. Puthuvelil, W. Lewis, K. Khokar, and Z. Erickson, “Robocap: Robotic classification and precision pouring of diverse liquids and granular media with capacitive sensing,” in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 19546–19553, 2025.
- [15] A. Bicchi, “Hands for dexterous manipulation and robust grasping: a difficult road toward simplicity,” *IEEE Transactions on Robotics and Automation*, vol. 16, no. 6, pp. 652–662, 2000.
- [16] S. Q. Liu, L. Z. Yañez, and E. H. Adelson, “Gelsight endoflex: A soft endoskeleton hand with continuous high-resolution tactile sensing,” 2023.
- [17] H. Wang, Z. Wang, G. Zhu, and S. Zhang, “Capacitive proximity skin for non-contact liquid state measurement of robot grippers with different structures,” *Sensors and Actuators A: Physical*, vol. 365, p. 114857, 2024.
- [18] Z. Zhao, W. Li, Y. Li, T. Liu, B. Li, M. Wang, K. Du, H. Liu, Y. Zhu, Q. Wang, K. Althoefer, and S.-C. Zhu, “Embedding high-resolution touch across robotic hands enables adaptive human-like grasping,” *Nature Machine Intelligence*, vol. 7, pp. 889–900, 2025.
- [19] S. E. Navarro, S. Mühlbacher-Karrer, H. Alagi, H. Zangl, K. Koyama, B. Hein, C. Duriez, and J. R. Smith, “Proximity perception in human-centered robotics: A survey on sensing systems and applications,” *IEEE Transactions on Robotics*, vol. 38, no. 3, pp. 1599–1620, 2022.
- [20] B. Kumar, G. Rajita, and N. Mandal, “A review on capacitive-type sensor for measurement of height of liquid level,” *Measurement and Control*, vol. 47, no. 7, pp. 219–224, 2014.
- [21] X. Liu, W. Yang, F. Meng, and T. Sun, “Material recognition using robotic hand with capacitive tactile sensor array and machine learning,” *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–9, 2024.
- [22] Y. Hu, H. Bai, R. Deng, and W. Yang, “Ect sensor for symmetrical robot gripper application,” *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–11, 2025.
- [23] R. Deng, Y. Hu, H. Bai, and W. Yang, “Material recognition with capacitive tactile sensor and robotic arm,” in *2025 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pp. 1–6, 2025.
- [24] H. Bai, J. Lu, R. Deng, Y. Hu, L. Cui, W. Yang, and H. Liu, “Sensor-driven filtering and contact planning for physically consistent tactile perception,” *IEEE Transactions on Instrumentation and Measurement*, vol. 75, pp. 1–12, 2026.
- [25] Y. Liu, Z. Shangguan, A. Tapus, S. Safin, F. D tienne, and E. Lecolinet, “Silicone-based haptic interfaces: Enhancing multimodal interactions through pneumatic tactile feedback,” in *2024 16th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pp. 140–143, IEEE, 2024.
- [26] B. Wu, T. Jiang, Z. Yu, Q. Zhou, J. Jiao, and M. L. Jin, “Proximity sensing electronic skin: Principles, characteristics, and applications,” *Advanced Science*, vol. 11, no. 13, p. 2308560, 2024.
- [27] Z. Shangguan, Y. Liu, L. Song, T. Li, and A. Tapus, “Using a pneumatic tactile steering wheel to enhance the multi-modal takeover request in smart vehicle,” in *International Conference on Social Robotics*, pp. 122–132, Springer, 2024.
- [28] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, “On layer normalization in the transformer architecture,” in *International conference on machine learning*, pp. 10524–10533, PMLR, 2020.