2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MFCC-Inspired Spectral Feature Extraction for Robust Touch Interaction in Social Robots

Ji Soo Kim[1], Sun Jun Hwang[1], Hyojin Kim[2], Dong Joon Hwang[1] and Hui Sung Lee[1]

***Abstract*— Touch is a fundamental modality for conveying emotions and intentions in Human–Robot Interaction. However, conventional approaches to touch pattern recognition often lack robustness to inter-user variability, whereas alternative solutions are frequently bulky or costly. This study proposes a novel feature extraction framework for touch pattern recognition, which adapts MFCC from speech processing to capacitive touch signals. The proposed method preserves the strengths of MFCC—dimensionality reduction and noise robustness—while addressing the physical differences between audio and touch signals by introducing a new frequency reference axis in place of the conventional Mel scale. To evaluate its effectiveness, a representative set of social touch patterns, including gestures traditionally difficult to classify, was defined and analyzed. The proposed framework ensures stable recognition across diverse users while reducing feature dimensionality for efficient operation in lightweight models. This efficiency highlights its suitability for real-time robotic interfaces**

I. INTRODUCTION

Achieving natural Human–Robot Interaction(HRI) requires the recognition of diverse interaction patterns, among which touch represents one of the most intuitive and fundamental modalities [1], [2]. Prior studies have demonstrated that touch effectively conveys intent and emotion even in the absence of visual cues [3], [4], underscoring its potential as a primary channel for social interaction in robotics.

Despite this promise, the role of touch in social robots has been limited. Capturing rich tactile patterns typically demands array-type or multimodal sensors, which introduce significant hardware cost and design complexity. Moreover, tactile sensing is inherently sensitive to inter-user variability and environmental factors, thereby reducing reproducibility and constraining the effectiveness of conventional machine learning approaches [5], [6]. Consequently, many robotic systems have downplayed touch as a principal modality due to both technical and practical challenges [7], [8].

In this work, we investigate capacitive touch sensors, which offer an attractive balance of simplicity and low cost. However, as with other tactile modalities, they are subject to substantial variability, which undermines reproducibility and poses a major barrier to robust classification [9]. To address this challenge, we leverage Mel-Frequency Cepstral Coefficients(MFCC), a frequency-domain feature extraction technique widely adopted in speech recognition [10]. Speech signals exhibit considerable variation across speakers and conditions, yet humans consistently achieve reliable recognition. MFCC-based approaches emulate this robustness and have achieved state-of-the-art performance in speech processing. Moreover, prior work has shown that applying MFCC to signals with low reproducibility improves classification performance, extending their utility beyond speech analysis [11], [12]. By analogy, we hypothesize that capacitive touch sensor signals, though variable across individuals and environments, exhibit consistent frequency-domain characteristics when generated with the same intent.

To evaluate this hypothesis, we collected data using monopolar capacitive touch sensors integrated into a custom-built social robot. We adapted the MFCC extraction pipeline from speech processing to the tactile domain, including the design of a novel frequency scale optimized for capacitive signals. The resulting features were assessed using lightweight machine learning models suited for embedded deployment.

The contributions of this work are threefold:(i) we introduce a frequency-domain framework for social touch interaction based on MFCC;(ii) we demonstrate robust recognition across diverse representative social touch patterns; and(iii) we validate the feasibility of deploying the proposed method for real-time HRI on resource-constrained robotic platforms.

II. RELATED WORK

A. Sensor Architectures for Touch Interaction

To enable robots to recognize diverse social touch patterns, prior studies have advanced beyond binary contact interfaces toward hardware architectures inspired by human skin. These approaches often utilize sensor designs that capture spatial features, thereby supporting the recognition of complex interaction patterns. Capacitive and piezoelectric sensors arranged in array configurations are among the most widely adopted solutions. While a single sensor provides only simple analog responses, arrays enable recognition across larger surface areas [13], [14]. This benefit, however, comes at the expense of high-dimensional inputs, which require substantial training data and computational resources.

In contrast to array-type sensors, some studies have introduced compliant three-dimensional structures embedded with pneumatic sensors to detect touch patterns with reduced computational requirements; however, such solutions often increase bulk and impose constraints on robot design [15] [16].

[1]Ji Soo Kim, Sun Jun Hwang, Dong Joon Hwang and Hui Sung Lee are with Design Department, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Korea {wisdomblink, sunjoon020, djhwang and huisung.lee}@unist.ac.kr

[2]Hyojin Kim with the Electrical Engineering Department, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Korea kim24@unist.ac.kr

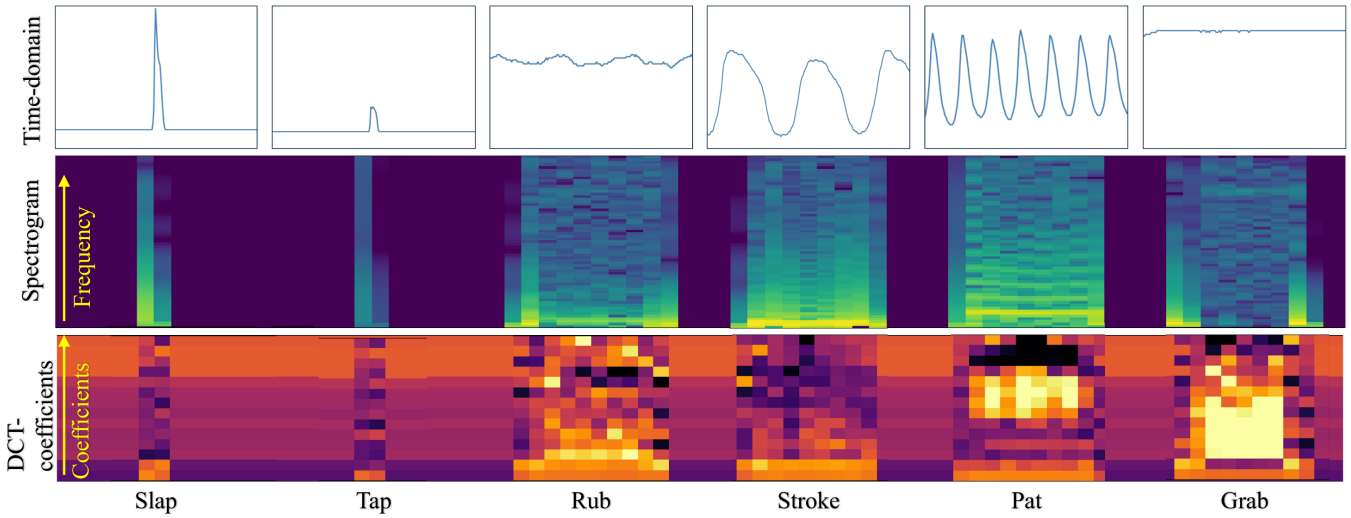

Fig. 1. Representative examples of sensor signals for each pattern class. The first row shows time-domain waveforms, the second row illustrates frequency-domain spectrograms, and the third row presents DCT coefficient spectrograms used as model inputs, normalized based on the scale of the collected data.

B. Touch pattern recognition

Touch pattern recognition has traditionally relied on waveform analysis of tactile sensor signals. Early studies extracted time-domain features such as peak points, signal duration, and force trajectories. With advances in hardware, spatial characteristics were also incorporated, leading to improvements in gesture classification performance.

A representative example is the use of array-type pressure sensors with a CNN-based recognition model operating without preprocessing. This work demonstrated that gestures could be classified from short signal segments while achieving higher accuracy than conventional feature-based classifiers [17]. Similarly, other studies showed that incorporating shear force in addition to normal force signals further enhanced recognition accuracy [18].

Despite progress driven by hardware improvements and the adoption of machine learning techniques, waveform-based approaches still exhibit inherent limitations. In practice, distinct gestures may produce highly similar signal waveforms, resulting in frequent misclassifications. Such ambiguity is further exacerbated by interpersonal variability in physical characteristics and habitual touch styles. Although research incorporating shear force has reported accuracy gains, the role of different frequency components—ranging from dominant low-frequency patterns to fine-grained friction-induced signals—has been only partially addressed.

In addition, model complexity and computational constraints remain critical considerations in embedded systems for robot control [19]. While omitting preprocessing may enhance real-time performance, in microcontroller environments equipped with DSP capabilities, lightweight preprocessing can reduce input dimensionality more efficiently than directly processing high-dimensional raw signals [20]. Furthermore, while LSTMs are generally unsuitable for deployment on microcontrollers due to their high computational cost, the use of CNNs becomes impractical primarily when dealing with high-dimensional inputs.

III. Feature Extraction

The core feature extraction method for touch pattern classification was designed by drawing inspiration from MFCC, a frequency-domain technique widely used in speech recognition. Fig. 1 illustrates the time-series signals and spectrograms of the touch patterns defined in this study. As shown in the first and second rows, clear spectral differences corresponding to each pattern can be visually observed, with dominant frequency components concentrated primarily in the lower bands. Based on these observations, we hypothesize that MFCC-based feature extraction can be effectively applied to touch sensor signals, provided that the method is adapted to account for the physical differences between touch signals and speech or audio data.

The data measured by the Touch Sensing Controller(TSC) on the robot's hardware control board were transmitted to the analysis software via a UART-to-USB interface. The sampling frequency f_s was defined based on the sampling period T_s. The resulting signal can be represented as a discrete-time sequence $x[n]$ that reflects variations in capacitance corresponding to touch events. In this study, the signal was analyzed using a framework inspired by the MFCC feature extraction process originally developed for speech signals Fig. 2. To accommodate the distinct characteristics of touch signals—such as their narrower frequency range and the distribution of dominant spectral components—the frequency scale was redesigned accordingly.

For frequency-domain analysis, the input sequence $x[n]$ was divided into fixed-length frames of size N. To compensate for temporal non-stationarity and to ensure sufficient frequency resolution, the hop size was set to half of the frame length, resulting in a 50% overlap between adjacent frames. A Hann window $w[n]$ was applied to each frame to mitigate spectral leakage caused by discontinuities at the

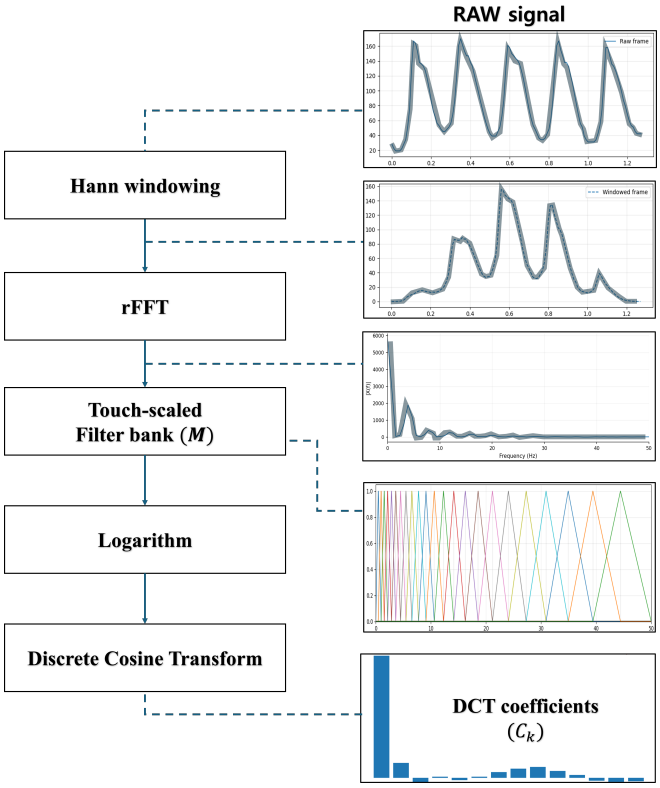

Fig. 2. Feature extraction pipeline. The raw sensor signal is windowed and transformed by rFFT, processed through M triangular filter banks, log-compressed, and finally decorrelated by DCT to yield K-dimensional feature vectors.

frame boundaries. The observed signal for one frame is thus expressed as follows, serving as the basic unit for subsequent frequency analysis:

$$x_w[n] = x[n] \cdot w[n], \quad 0 \le n < N \tag{1}$$

To analyze the frame-based and windowed signals in the frequency domain, the rFFT(real-valued Fast Fourier Transform) was applied. The spectrum of a single frame after windowing is defined as follows:

$$X[k] = \sum_{n=0}^{N-1} x_\omega[n] e^{-j2\pi kn/N}, \quad 0 \le k \le \frac{N}{2} \tag{2}$$

$X[k]$ corresponds to the frequency defined as $f_k = \frac{k}{N} f_s$, and the spectrum magnitude $|X[k]|$ was used as the input for feature extraction. In the conventional MFCC method, the filter bank is arranged according to the Mel scale, an empirically derived frequency scale reflecting the non-linear sensitivity of the human auditory system across the audible range. The Mel scale is defined as follows:

$$m(f) = A \log\left(1 + \frac{f}{f_0}\right) \tag{3}$$

On the Mel scale, the low-frequency region is represented with finer resolution, whereas the high-frequency region is compressed. This property provides an advantage in modeling the perceptual characteristics of speech signals. The parameter f_0 serves as a transition (knee) point, indicating where the scale shifts from an approximately linear resolution at low frequencies to a logarithmically compressed resolution at higher frequencies. However, because the Mel scale was specifically designed to emulate the human auditory system, it cannot be assumed to be optimal for the touch sensor signals considered in this study. To address this, we define a new reference scale adapted to the actual frequency distribution of the sensor signals.

For this purpose, Welch's method was employed to estimate the PSD(Power Spectral Density) of the signal, from which the cumulative energy $E(f)$ was derived [21]. Based on the estimated PSD, the cumulative energy across the frequency axis is computed as follows:

$$\begin{aligned} E(f) &= \sum_{k=0}^{j} P_{xx}(f_k)\,\Delta f, \\ C(f_j) &= \frac{E(f_j)}{E(f_{\max})}, \qquad f_0 \quad = C^{-1}(\theta) \end{aligned} \tag{4}$$

f_k denotes the frequency corresponding to the k-th FFT bin, which can be derived from the formulation in Eq. (2), with bin spacing $\Delta f = f_s/N$. The function $C(f_j)$ represents the normalized cumulative energy along the frequency axis, bounded between 0 and 1. Accordingly, $\theta \in (0, 1)$ denotes the target cumulative energy ratio, and f_0 is defined as the frequency at which the cumulative energy reaches this ratio. The resulting reference frequency f_0, incorporated into Eq.(3), enables effective separation between the low-frequency region—where fine resolution is required—and the high-frequency region—where compressed analysis is sufficient—without reliance on the traditional Mel scale. Based on this reference, the filter bank is rearranged such that each filter $H_m(f)$ is defined in a triangular form. Specifically, each filter is constructed from the center frequency f_m and the adjacent frequencies f_{m-1} and f_{m+1}, as expressed below:

$$H_m(f) = \begin{cases} 0, & f < f_{m-1} \\ \dfrac{f - f_{m-1}}{f_m - f_{m-1}}, & f_{m-1} \le f \le f_m \\ \dfrac{f_{m+1} - f}{f_{m+1} - f_m}, & f_m \le f \le f_{m+1} \\ 0, & f > f_{m+1} \end{cases} \tag{5}$$

Based on these functions, the energy within each frequency band is aggregated to derive the fundamental components of the final feature vector. Specifically, by applying the filter response as a weighted sum over the PSD $P_{xx}(f)$ of the framed signal, the output energy of the m-th filter, E_m, is computed as follows:

$$E_m = \sum_f P_{xx}(f) \cdot H_m(f) \tag{6}$$

E_m represents the total energy of the frequency components within the m-th band. Repeating this calculation across all bands yields the vector $E = [E_1, E_2, E_3, \ldots, E_M]$. These values are subsequently transformed using a logarithmic operation to compensate for scale differences across bands and to produce a representation suitable for the Discrete Cosine Transform(DCT). When the number of filters is set to M and the number of feature coefficients to K, the k-th feature coefficient is defined as follows:

$$c_k = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} \tilde{E}_m \cos\left(\frac{\pi k}{M}\left(m - \frac{1}{2}\right)\right) \tag{7}$$

where $0 \leq k \leq K-1$ and $\tilde{E}_m = \log E_m$.

The coefficients obtained through this transformation, $c = [c_0, c_1, c_2, \ldots, c_{K-1}]$, provide a compact representation of the frequency structure of the original signal while reducing inter-coefficient correlation, thereby forming an effective feature vector. While inheriting the strengths of MFCC in capturing essential spectral characteristics, the proposed approach adapts the concentration of information in the lower-order coefficients to the properties of touch sensor signals. In this way, the method emphasizes the most informative components in a manner tailored to the sensing modality, rather than relying solely on auditory-inspired scaling.

IV. Implementation

A. Hardware setup and Data acquisition

This study's touch interaction experiment was conducted using the PO-ME social robot, designed to support children's reading activities [22], [23]. Fig. 3 illustrates the experimental hardware setup. The left shows the PO-ME social robot used in this study, while the top-right panel depicts the touch electrode attached to the robot's forehead using a thin copper film. The bottom-right schematic shows how the electrode was connected to the TSC port of the STM32F072 MCU through a 1 kΩ resistor and a 47 nF capacitor. The

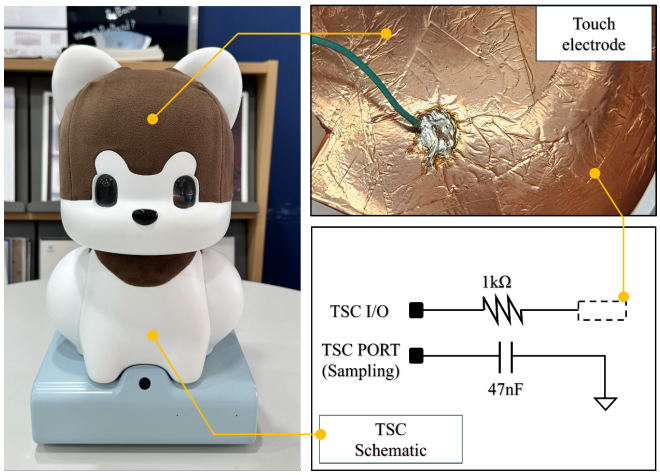

Fig. 3. Experimental hardware setup: the PO-ME robot (left), the forehead touch electrode (top-right), and its connection to the STM32F072 TSC (bottom-right).

TABLE I
DEFINITIONS OF SIX SOCIAL TOUCH PATTERNS EMPLOYED FOR INTERACTION EXPERIMENTS.

Gesture	Description
Slap	Quick, sharp strike with the open hand.
Tap	Light, brief touch with fingertips.
Rub	Firm, repetitive motion.
Stroke	Gentle, continuous caress along the surface.
Pat	Soft, quick tap with the palm.
Grab	Forceful grasp with the hand.

experimental data were acquired through the TSC of the STM32F072 MCU on the custom PO-ME control board and transferred via a USB–UART bridge. The F072-based TSC was configured according to the official reference manual [24], and the sampling frequency was set to 100 Hz in accordance with TI's application note [25] to ensure reliable detection of tap and stroke patterns.

In this mode, the TSC-IOGxCR counter decreases monotonically with increasing touch intensity. For consistency, the raw values were reverse-linearly normalized relative to the no-touch baseline and the stable-contact threshold, yielding an 8-bit scale (0–255). This normalization improved UART transfer efficiency and ensured consistency in subsequent signal processing.

B. Touch pattern definition

Based on observations of child–robot interactions, six touch patterns frequently directed toward the robot's head were selected: Slap, Tap, Rub, Stroke, Pat, and Grab. The naming and definitions of these gestures followed Yohanan and MacLean's social touch dictionary [1], with Slap constrained to a light contact to preserve the robot's durability. These patterns are both representative of natural interactions and challenging to discriminate under conventional waveform-based sensor analysis [17]. Accordingly, this study employed these six patterns (Table I) as the experimental dataset for evaluating the proposed feature extraction method.

C. Data annotation

Labels were assigned on the same framing used for feature extraction (frames of $N = 128$ with 50% overlap; see Eqs. (1)–(2)). For short and transient patterns such as *taps*, labels were assigned at points showing abrupt spectral changes, whereas sustained patterns such as *strokes* were annotated over their full duration. Because features were extracted on a frame basis, frame-level labels were determined using a majority voting rule: if a single class occupied at least 80% of the samples within a frame, that frame was assigned to the corresponding class. Frames that did not meet this criterion were discarded to avoid ambiguous training examples. This empirical threshold was selected to reduce labeling noise near event boundaries and to enhance the robustness of subsequent classification.

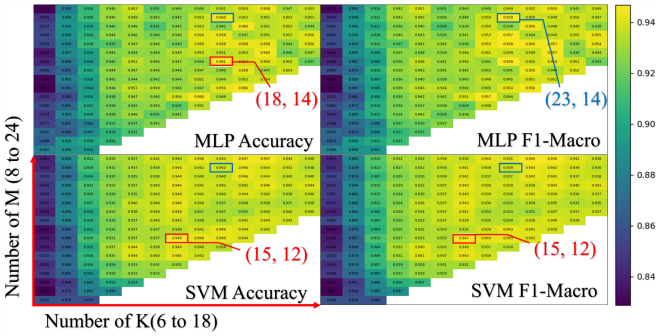

Fig. 4. Grid search results of parameter optimization with respect to the number of filter banks M and DCT coefficients K. The plots compare MLP and SVM in terms of Accuracy and F1-Macro, with the best-performing configurations highlighted.

D. Experimental Protocol

A total of 15 participants (8 male, 7 female) were recruited, comprising 14 Koreans and 1 Pakistani. Only aggregate demographic data were retained to ensure privacy.

Participants performed the six touch interactions defined in the previous section on the forehead of the robot. To minimize variability in how participants interpreted the definitions of each pattern (e.g., confusing *Rub* with *Stroke*), the experimenter provided a brief demonstration prior to data collection. Execution, however, was left to each participant's natural style as long as it conformed to the dictionary definitions. Sensor inputs were continuously monitored via a serial plotter to prevent missed detections or erroneous samples. Data for each pattern were collected as follows:

- Slap and Tap: 50 repetitions each, with at least a 2-second interval.
- Rub, Stroke, Pat, and Grab: 10-second trials, repeated 5 times each (no interval constraints).
- Non-touch intervals were automatically labeled as untouched.

In addition to the primary data collection task, a separate dataset was gathered for independent evaluation, consisting of five repetitions of *Slap* and *Tap*, and a single 10-second trial of the remaining patterns per participant.

E. Parameter optimization and Model training

The final training dataset comprised 5,616 frames, which were used as an *internal holdout*. An additional dataset collected during the experiment was strictly excluded from the training process and is hereafter defined as the *external dataset* for independent evaluation; this dataset consisted of 974 frames. For model evaluation, we primarily employed a MLP, given its favorable trade-off between accuracy and computational efficiency [26]–[29]. A SVM was additionally included as a baseline model to monitor potential overfitting.

The proposed pipeline was parameterized by the reference frequency f_0, the number of filter banks M, and DCT coefficients K. Since no standard values exist for capacitive touch signals, we set θ in Eq. (4) to 0.8 to emphasize low-frequency energy. Optimal parameter values for (M, K) were determined through grid search, as reported in the next section. To further assess subject-independent generalization, we also conducted LOSO CV(Leave-One-Subject-Out Cross-Validation) across all 15 participants, in which data from one participant were held out for testing while the remaining participants' data were used for training.

TABLE II

PERFORMANCE OF MLP AND SVM WITH OPTIMIZED PARAMETERS ($f_0 = 3.91, M = 23, K = 14$), EVALUATED ON INTERNAL HOLDOUT AND INDEPENDENT EXTERNAL DATASET.

Model	Dataset	Accuracy	F1-Macro	95% CI
MLP	Internal	0.979	0.981	[0.969, 0.990]
MLP	External	0.940	0.910	[0.887, 0.931]
SVM	Internal	0.942	0.939	[0.916, 0.957]
SVM	External	0.941	0.906	[0.881, 0.927]

TABLE III

PERFORMANCE OF BASELINE MODELS WITH RAW AND RFFT INPUTS. RESULTS ARE SHOWN FOR INTERNAL HOLDOUT AND EXTERNAL DATASETS (ACCURACY AND F1-MACRO).

Model	Internal holdout		External	
	Accuracy	F1-Macro	Accuracy	F1-Macro
Raw data input				
1-D CNN	0.920	0.926	0.500	0.698
LSTM	0.874	0.883	0.895	0.859
MLP	0.938	0.942	0.504	0.704
SVM	0.915	0.917	0.482	0.670
rFFT input				
1-D CNN	0.832	0.810	0.412	0.562
LSTM	0.585	0.509	0.184	0.240
MLP	0.943	0.948	0.500	0.700
SVM	0.941	0.950	0.500	0.698

V. RESULT

Fig. 4 shows the grid search results for both MLP and SVM. The SVM achieved its best performance at $(M, K) = (15, 12)$, whereas the MLP reached its highest Accuracy at $(18, 14)$ and its highest F1-Macro at $(23, 14)$. Within this parameter region, the MLP achieved an Accuracy of 0.96, while the SVM reached 0.942 Accuracy and 0.939 F1, indicating that both models operated close to their peak performance and that the region provided stable results. Considering deployment requirements and class imbalance, we selected $(M, K) = (23, 14)$ with $f_0 = 3.91$ Hz as the optimal configuration.

As summarized in Table II, both models achieved high accuracy on internal and external datasets, with the MLP

TABLE IV

SUMMARY STATISTICS OF MLP PERFORMANCE UNDER LOSO CV (15 FOLDS). REPORTED VALUES ARE MEAN, STANDARD DEVIATION (SD), MINIMUM, AND MAXIMUM ACROSS PARTICIPANTS.

Metric	Mean	SD	Min	Max
Accuracy	0.930	0.036	0.860	0.980
Balanced Acc	0.926	0.045	0.831	0.978
F1-Macro	0.925	0.037	0.843	0.977
95% CI (low)	0.891	0.054	0.761	0.955
95% CI (high)	0.945	0.042	0.842	0.993

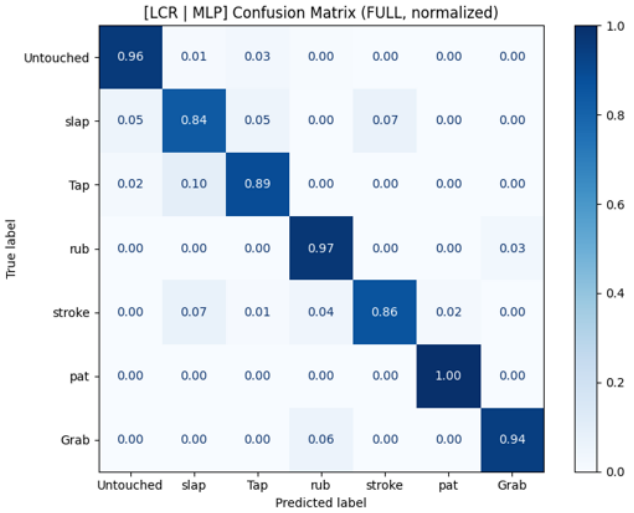

Fig. 5. Normalized confusion matrix of the MLP model evaluated on an independent external dataset (not used during training). The results confirm stable classification performance and demonstrate the model's generalization capability beyond the training data.

maintaining slightly better generalization. Notably, the MLP required only 1,696 MACs and 11 KB of memory per frame, compared to 14,420 MACs and 226 KB for the SVM.

To quantify practical execution cost, on-device inference was evaluated using an STM32H755 microcontroller with DSP support. The deployed model was the lightweight MLP used in the experiments and was implemented using STMicroelectronics' X-Cube-AI framework [30]. End-to-end processing, including feature extraction and model inference, required 164 μs per frame, while the memory footprint of the processing task was measured to be 624 Bytes. The execution time was measured using GPIO-based timing with an oscilloscope.

Table III presents the results of baseline models trained with raw data and rFFT inputs. All four models (1-D CNN, LSTM, MLP, and SVM) were evaluated under identical conditions, and results are reported for both internal holdout and external datasets. The table indicates that while each baseline achieved reasonable performance on the internal dataset, their accuracy and F1-Macro values declined on the external dataset. Detailed model configurations were aligned in parameter size ($\approx$ 1.5k) to enable fair comparison.

Finally, Table IV summarizes the overall performance of the MLP under LOSO CV across 15 participants. The average F1-Macro was 0.925 ($\pm$0.037), with values ranging from 0.843 to 0.977. Accuracy and Balanced Accuracy showed similar trends (means of 0.930 and 0.926, respectively), while 95% confidence intervals remained narrow overall.

VI. Discussion

The proposed feature extraction pipeline, inspired by MFCC in speech recognition, was developed under the hypothesis that touch interactions share common frequency components across individuals. Experimental results confirmed that distinct touch patterns exhibit separable spectral

TABLE V
Performance metrics (Precision, Recall, F1) of the MLP model on the independent external dataset

Pattern	Precision	Recall	F1
Untouched	1.00	0.96	0.98
Slap	0.75	0.84	0.79
Tap	0.78	0.89	0.83
Rub	0.89	0.97	0.93
Stroke	0.97	0.86	0.91
Pat	0.96	1.00	0.98
Grab	0.96	0.94	0.95

characteristics (Fig. 1). Spectral estimation using the Welch method further showed that dominant frequency components concentrate in the lower bands, supporting the adaptation of the MFCC framework to capacitive touch signals.

Both the MLP and the baseline SVM achieved high performance on the internal holdout and the independent external dataset. Each class attained an accuracy between 0.84 and 0.97 on the external dataset (Fig. 5), with only limited degradation compared to the internal results. These findings indicate that the extracted features generalize effectively across patterns while keeping input dimensionality low. The MLP required only 1,696 MACs and 11 KB of memory, demonstrating suitability for deployment in resource-constrained environments, as corroborated by the on-device inference evaluation.

Baseline models trained directly on raw or rFFT inputs were also evaluated (Table III). While these models produced competitive results on the internal dataset, their performance declined substantially on the external dataset. This outcome suggests that end-to-end training on high-dimensional inputs failed to capture generalizable features of touch interactions. In addition, the large input dimensionality resulted in higher parameter counts or computational costs, further limiting their applicability to real-time embedded systems. For fair comparison, all baseline models were constrained to a parameter budget comparable to the proposed MLP, which may have further limited their ability to compensate for the lack of domain-specific features. These contrasts highlight the

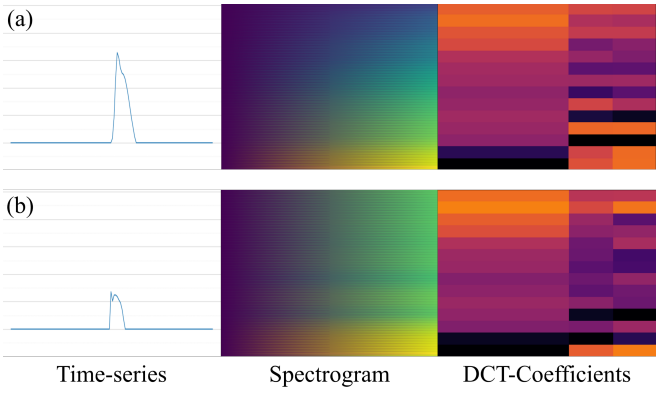

Fig. 6. Comparison of touch gesture patterns over a 1.28s window. Time-series signals, spectrograms, and DCT coefficients are shown for (a) Slap and (b) Tap gestures.

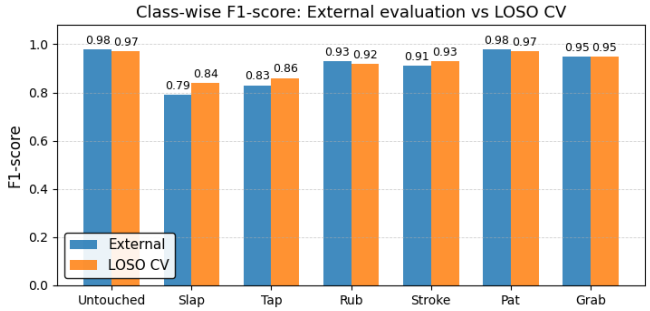

Fig. 7. Class-wise F1 scores of the MLP on the External dataset and LOSO CV. Pat, Grab, and Untouched maintain consistently high performance, while transient gestures such as Slap and Tap remain relatively weaker across both settings.

necessity of compact, domain-tailored features for robust and efficient social touch recognition.

Table V summarizes the class-wise results on the external dataset. Pat, Rub, Stroke, and Grab achieved consistently high accuracy, whereas Tap and Slap showed relatively lower performance. This behavior arises from the intrinsic similarity of impulsive contact signals in the cepstral feature space. As illustrated in Fig. 6, the two patterns are primarily distinguished by amplitude while exhibiting similar spectral shapes. Furthermore, since the impulse duration is much shorter than the analysis window (1.28 s), log compression and temporal averaging cause their spectral statistics to converge, leading to reduced performance in spectrum-based classification.

LOSO CV further evaluated subject-independent generalization. The average F1-Macro remained high (0.925), but some folds dropped to 0.843, suggesting sensitivity to subject-specific factors such as touch intensity or consistency. Despite this variability, the narrow confidence intervals highlight the robustness of the pipeline. Future work will investigate the sources of subject variability and explore strategies to improve consistency across diverse users.

As illustrated in Fig. 7, the class-wise F1-scores under both External and LOSO evaluations reveal stable performance for Pat, Grab, and Untouched (F1 $\geq$ 0.95). Rub and Stroke also remained robust, while Slap and Tap were comparatively weaker (0.79–0.86). These transient gestures are highly sensitive to execution style, making them more prone to misclassification. Overall, the analysis shows that the proposed approach generalizes well for sustained interactions, while short-lived events remain challenging, motivating future work on temporal modeling and adaptive calibration.

VII. Conclusion

This study presented a feature extraction framework for social touch recognition by adapting the MFCC technique to capacitive touch signals. The proposed method employed frequency scaling derived from sensor data, enabling the capture of dominant low-frequency components while reducing input dimensionality and preserving discriminative power.

Experimental results demonstrated that the extracted features generalized well across internal and external datasets, with lightweight models such as MLPs achieving high performance under stringent resource constraints. Although transient gestures (e.g., Slap and Tap) remained more challenging due to their short duration and variability, sustained patterns such as Rub and Grab were consistently recognized with high accuracy.

These findings support the practicality of the proposed approach for real-time deployment on microcontroller-based Edge-AI platforms. Future work will investigate hybrid approaches that integrate temporal dynamics and extend evaluation to more diverse user groups, with the aim of further improving robustness and applicability in social robotics.

Beyond its technical contributions, this work highlights the potential of low-cost capacitive sensing to support more natural and expressive human–robot interaction. By lowering computational and hardware barriers, the proposed framework contributes to scalable deployment of social touch interfaces in everyday robotic systems and assistive technologies.

References

[1] S. Yohanan and K. E. MacLean, "The role of affective touch in human-robot interaction: Human intent and expectations in touching the haptic creature," *International Journal of Social Robotics*, vol. 4, no. 2, pp. 163–180, 2012. [Online]. Available: https://doi.org/10.1007/s12369-011-0126-7

[2] Y.-K. Li, Q.-H. Meng, T.-H. Yang, Y.-X. Wang, and H.-R. Hou, "Touch gesture and emotion recognition using decomposed spatiotemporal convolutions," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–9, 2022. [Online]. Available: https://doi.org/10.1109/TIM.2022.3147338

[3] M. J. Hertenstein, J. M. Verkamp, A. M. Kerestes, and R. M. Holmes, "The communicative functions of touch in humans, nonhuman primates, and rats: a review and synthesis of the empirical research," *Genetic, Social, and General Psychology Monographs*, vol. 132, no. 1, pp. 5–94, Feb. 2006. [Online]. Available: https://doi.org/10.3200/mono.132.1.5-94

[4] I. Morrison, L. S. Löken, and H. Olausson, "The skin as a social organ," *Experimental Brain Research*, vol. 204, no. 3, pp. 305–314, Jul. 2010. [Online]. Available: https://doi.org/10.1007/s00221-009-2007-y

[5] V. Gonzalez-Pacheco, A. Ramey, F. Alonso-Martin *et al.*, "Maggie: A social robot as a gaming platform," *International Journal of Social Robotics*, vol. 3, no. 4, pp. 371–381, 2011. [Online]. Available: https://doi.org/10.1007/s12369-011-0109-8

[6] S. Domínguez-Gimeno, R. Igual-Catalán, and I. Plaza-García, "Sensor arrays: A comprehensive systematic review," *Sensors*, vol. 25, no. 16, p. 5089, 2025. [Online]. Available: https://doi.org/10.3390/s25165089

[7] M. J. Hertenstein, R. Holmes, M. McCullough, and D. Keltner, "The communication of emotion via touch," *Emotion*, vol. 9, no. 4, pp. 566–573, Aug. 2009. [Online]. Available: https://doi.org/10.1037/a0016108

[8] M. J. Hertenstein, D. Keltner, B. App, B. A. Bulleit, and A. R. Jaskolka, "Touch communicates distinct emotions," *Emotion*, vol. 6, no. 3, pp. 528–533, Aug. 2006. [Online]. Available: https://doi.org/10.1037/1528-3542.6.3.528

[9] A. Hoffmann, A. Poeppel, A. Schierl, and W. Reif, "Environment-aware proximity detection with capacitive sensors for human-robot-interaction," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, Korea (South), 2016, pp. 145–150. [Online]. Available: https://doi.org/10.1109/IROS.2016.7759047

[10] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980. [Online]. Available: https://doi.org/10.1109/TASSP.1980.1163420

[11] R. San-Segundo, J. M. Montero, R. Barra-Chicote, F. Fernández, and J. M. Pardo, "Feature extraction from smartphone inertial signals for human activity segmentation," *Signal Processing*, vol. 120, pp. 359–372, 2016. [Online]. Available: https://doi.org/10.1016/j.sigpro.2015.09.029

[12] T. Tsuji, K. Sato, and S. Sakaino, "Contact feature recognition based on mfcc of force signals," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5153–5158, Jul. 2021. [Online]. Available: https://doi.org/10.1109/LRA.2021.3072035

[13] T. Minato, Y. Yoshikawa, T. Noda, S. Ikemoto, H. Ishiguro, and M. Asada, "Cb2: A child robot with biomimetic body for cognitive developmental robotics," in *2007 7th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, Pittsburgh, PA, USA, 2007, pp. 557–562. [Online]. Available: https://doi.org/10.1109/ICHR.2007.4813926

[14] N. Mitsunaga, T. Miyashita, H. Ishiguro, K. Kogure, and N. Hagita, "Robovie-iv: A communication robot interacting with people daily in an office," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China, 2006, pp. 5066–5072. [Online]. Available: https://doi.org/10.1109/IROS.2006.282594

[15] K. Park *et al.*, "A biomimetic elastomeric robot skin using electrical impedance and acoustic tomography for tactile sensing," *Science Robotics*, vol. 7, p. eabm7187, 2022. [Online]. Available: https://doi.org/10.1126/scirobotics.abm7187

[16] K. Park, K. Shin, S. Yamsani, K. Gim, and J. Kim, "Low-cost and easy-to-build soft robotic skin for safe and contact-rich human–robot collaboration," *IEEE Transactions on Robotics*, vol. 40, pp. 2327–2338, 2024. [Online]. Available: https://doi.org/10.1109/TRO.2024.3378174

[17] S. Albawi, O. Bayat, S. Al-Azawi, and O. N. Ucan, "Social touch gesture recognition using convolutional neural network," *Computational Intelligence and Neuroscience*, vol. 2018, p. Article ID 6973103, 2018. [Online]. Available: https://doi.org/10.1155/2018/6973103

[18] H. Choi *et al.*, "Deep learning classification of touch gestures using distributed normal and shear force," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Kyoto, Japan, 2022, pp. 3659–3665. [Online]. Available: https://doi.org/10.1109/IROS47612.2022.9981457

[19] S. S. Saha, S. S. Sandha, and M. Srivastava, "Machine learning for microcontroller-class hardware: A review," *IEEE Sensors Journal*, vol. 22, no. 22, pp. 21 362–21 390, Nov. 2022. [Online]. Available: https://doi.org/10.1109/JSEN.2022.3210773

[20] A. Elsts and R. McConville, "Are microcontrollers ready for deep learning-based human activity recognition?" *Electronics*, vol. 10, no. 21, p. 2640, 2021. [Online]. Available: https://doi.org/10.3390/electronics10212640

[21] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, Jun. 1967. [Online]. Available: https://doi.org/10.1109/TAU.1967.1161901

[22] J. S. Kim, S. J. Hwang, M. J. Sung, Y. Kwon, and H. S. Lee, "System architecture and hardware design of a dog-type robot for children reading activities," in *2024 24th International Conference on Control, Automation and Systems (ICCAS)*, Jeju, Korea, Republic of, 2024, pp. 989–994. [Online]. Available: https://doi.org/10.23919/ICCAS63016.2024.10773382

[23] Y. Kwon, S. Jeong, H. Park, and H. S. Lee, "Design of a dog-type social robot to support children's reading activities and development of a touch sensor module for users' touch interaction," in *2023 23rd International Conference on Control, Automation and Systems (ICCAS)*, Yeosu, Korea, Republic of, 2023, pp. 1683–1688. [Online]. Available: https://doi.org/10.23919/ICCAS59377.2023.10316982

[24] *RM0091 Reference Manual: STM32F0x1/STM32F0x2/STM32F0x8 advanced Arm-based 32-bit MCUs*, Rev. 10 ed., STMicroelectronics, May 2022. [Online]. Available: https://www.st.com/resource/en/reference_manual/rm0091-stm32f0x1stm32f0x2stm32f0x8-advanced-arm-based-32bit-mcus-stmicroelectronics.pdf

[25] D. Lehman, "Capacitive touch gesture software and tuning," Texas Instruments, Tech. Rep. SLAA877, Dec. 2018, [Online]. Available: https://www.ti.com/lit/an/slaa877/slaa877.pdf.

[26] F. Noble, M. Xu, and F. Alam, "Static hand gesture recognition using capacitive sensing and machine learning," *Sensors*, vol. 23, no. 7, p. 3419, 2023. [Online]. Available: https://doi.org/10.3390/s23073419

[27] Q. Zhao, F. Wang, W. Wang *et al.*, "Research on intrusion detection model based on improved mlp algorithm," *Scientific Reports*, vol. 15, p. 5159, 2025, published 12 Feb. 2025. [Online]. Available: https://doi.org/10.1038/s41598-025-89798-0

[28] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Extracting deep neural network bottleneck features using low-rank matrix factorization," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 185–189. [Online]. Available: https://doi.org/10.1109/ICASSP.2014.6853583

[29] A. Dharmawan, R. E. Masithoh, and H. Z. Amanah, "Development of pca-mlp model based on visible and shortwave near infrared spectroscopy for authenticating arabica coffee origins," *Foods*, vol. 12, no. 11, p. 2112, 2023. [Online]. Available: https://doi.org/10.3390/foods12112112

[30] *UM2526 User Manual: Getting started with X-CUBE-AI Expansion Package for Artificial Intelligence (AI)*, Rev. 7 ed., STMicroelectronics, Feb. 2023. [Online]. Available: https://www.st.com/resource/en/user_manual/um2526-getting-started-with-xcubeai-expansion-package-for-artificial-intelligence-ai-stmicroelectronics.pdf