

# Tap2Stop: Pedestrian-Activated Emergency Stop for Autonomous Vehicles

Xinyue Gui  
The University of Tokyo  
Tokyo, Japan  
xinyuegui@gmail.com

Ding Xia  
The University of Tokyo  
Tokyo, Japan  
dingxia1995@gmail.com

Yuan Li  
Graduate School of Information  
Science and Technology  
The University of Tokyo  
Tokyo, Japan  
liyuan@g.ecc.u-tokyo.ac.jp

Mark Colley  
UCL Interaction Centre  
London, United Kingdom  
m.colley@ucl.ac.uk

Chia-Ming Chang  
National Taiwan University of Arts  
Taipei, Taiwan  
iamchiaming@gmail.com

Stela Hanbyeol Seo  
Social Informatics  
Kyoto University  
Kyoto, Kyoto, Japan  
stela.seo@i.kyoto-u.ac.jp

Manabu Tsukada  
The University of Tokyo  
Tokyo, Japan  
mtsukada@g.ecc.u-tokyo.ac.jp

Takeo Igarashi  
The University of Tokyo  
Tokyo, Japan  
takeo@acm.org

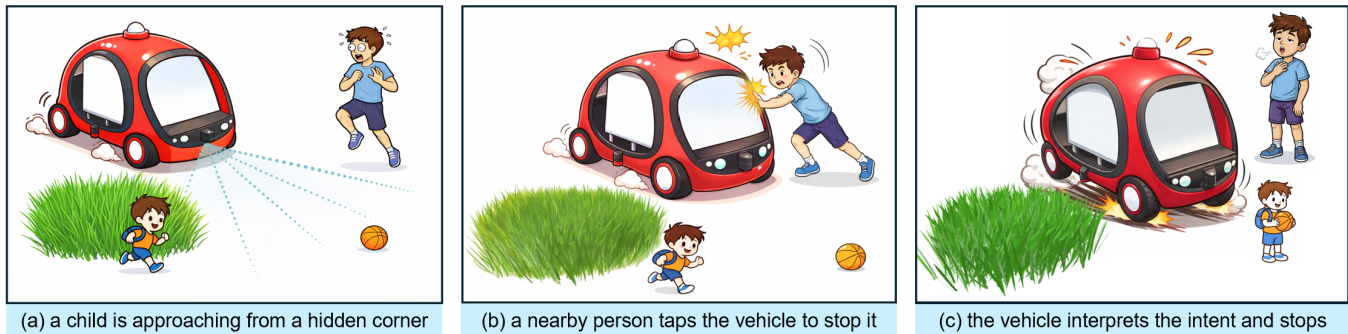


**Figure 1: Overview of *Tap2Stop*, an intuitive, human-centered method that enables pedestrians to actively stop autonomous vehicles by tapping. a) a child suddenly enters the path of an autonomous vehicle driving slowly in a complex environment (i.e., the grass blocks the detection of the moving child), b) a nearby adult taps the vehicle to stop it, and c) the vehicle interprets the intent and stops. This method empowers pedestrians to become active roles in AV safety.**

## Abstract

Autonomous vehicles (AVs) remain imperfect in close-range interactions, where sensor blind spots and unpredictable pedestrian behavior lead to safety-critical failures. Existing emergency-stop mechanisms are typically limited to remote operators or passengers, leaving pedestrians unable to intervene. We propose *Tap2Stop*, an externally accessible mechanism that enables pedestrians to “actively” stop vehicles by tapping on the surface. A formative study (23 pedestrians, 8 AV developers) reveals a gap between pedestrians’ desire for intervention and developers’ concerns about misuse.

Guided by these insights, we develop a tapping-based system integrating two classifiers (tapping detection and intent classification) with a gating and accumulation mechanism to balance robustness and reaction time. Study 1 constructs a dataset across multiple conditions to train the classifiers and optimize parameters. Study 2 evaluates the system in a field study (N=18), showing promising performance (100% detection rate, 24% false positive rate), user acceptance, and providing insights for future bidirectional human-vehicle interaction.



This work is licensed under a Creative Commons Attribution 4.0 International License.  
UIST '26, Detroit, MI, USA  
© 2026 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-2856-3/26/11  
<https://doi.org/10.1145/3830398.3830497>

## CCS Concepts

• Human-centered computing → Interaction design process and methods.

## Keywords

pedestrian-vehicle interaction, pedestrian-activated emergency stop, tapping-based intent recognition, autonomous vehicles

### ACM Reference Format:

Xinyue Gui, Ding Xia, Yuan Li, Mark Colley, Chia-Ming Chang, Stela Hanbyeol Seo, Manabu Tsukada, and Takeo Igarashi. 2026. Tap2Stop: Pedestrian-Activated Emergency Stop for Autonomous Vehicles. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26)*, November 02–05, 2026, Detroit, MI, USA. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3830398.3830497>

## 1 Introduction

Accidents in real-world deployments have led to recalls of *Full Self-Driving* automated vehicles (AVs) [42]. For example, a Cruise AV in San Francisco failed to detect a pedestrian, resulting in the person being trapped under the vehicle, and dragged over 20 feet [4]. Another case (Figure 1) is about a sudden appearance from a blind spot, where an oncoming vehicle cannot see it in time to react [2]. These incidents highlight that AVs remain imperfect. They suffer from detection failures in near-field blind zones and struggle to anticipate and respond to unpredictable pedestrian behaviors. While such failures are rare, they raise an important question: if pedestrians recognize immediate danger when an AV fails to react, or before it reacts, should they have a way to actively communicate “stop”?

To explore this challenge, pedestrians can be active participants who can interrupt the vehicle in an emergency situation (e.g., the adult in Figure 1). Even though AVs include emergency stop mechanisms, such as remote-stop devices [39] or internal controls (e.g., driver-assistance tools [49] and emergency stop buttons [36]), these have limitations and do not include the pedestrians. For example, remote-stop devices are not easily accessible to general public. Passengers may not be able to perceive hazards outside of vehicle [42], or the AV may not include passengers. As a result, **pedestrians play only a passive role** in safety-critical situations, leaving them without a means to intervene AVs, even if they can foresee impending danger or are at direct risk.

We propose *Tap2Stop*, an externally accessible emergency mechanism that enables nearby pedestrians to **actively take an action**. The vehicle surface serves as a touchable interface, allowing pedestrians to trigger an emergency stop by **tapping**. By interpreting tapping patterns via multi-feature classification, *Tap2Stop* infers pedestrians’ requests. We consider this approach intuitive, as it is similar to behaviors people already use to notify human drivers, extending to AVs. The design targets close-range, low-speed scenarios, where a vulnerable person or object may fall into sensor blind spots [8, 46] and primary systems (e.g., perception or communication) can fail. In such situations, providing a simple, last-resort intervention can be life-saving. That is, our goal is not to replace existing AV perception or emergency braking systems, but to explore whether a fallback intervention channel could provide additional value in such perception-failure scenarios. In this work, we present an initial exploration of this interaction concept through a prototype and a field study.

To operationalize the “tap” concept into a practical design, we conducted a formative study (questionnaires and interviews) with

pedestrians (N=23) and AV developers (N=8), revealing a gap between pedestrians’ desire for direct intervention and developers’ concerns about misuse. This study informed two core design requirements: intuitive operation under time pressure and robustness against accidental or malicious activation.

Based on the identified requirements, we implement a tapping-based system using two coordinated classifiers: a tapping detector and an intent classifier that filters non-intentional and malicious inputs. To balance robustness and reaction time, we combine their outputs through a gating and accumulation mechanism that selectively validates signals and aggregates confidence over time.

We validate *Tap2Stop* through two studies. In **Study 1**, we collected a dataset of diverse tapping types from 23 participants to train and calibrate the classifiers. We then conducted a **classifier-centered analysis** to identify effective features, report accuracy, and determine key parameters for system optimization. In **Study 2**, we integrated the trained classifiers and calibrated parameters from Study 1 with a triggering mechanism into an initial prototype of *Tap2Stop*, which was evaluated through a field study with 18 participants. We then performed a **pedestrian-centered analysis** to understand how pedestrians perceive this method.

Results show that *Tap2Stop* showed promising performance in situ, both in recognizing intentional stop attempts and in filtering malicious tapping, while leaving room for further improvement in robustness. Questionnaire results indicate that the method is well accepted by pedestrians, leading to increased trust in AVs. We further identify three types of feedback that AVs should provide in response to pedestrian tapping (safety-related information, future action guidance, and accountability), which participants requested to support bidirectional interaction through visual or haptic cues.

### Contribution Statement [53]

- **Interaction Technique.** We introduce *Tap2Stop*, a **direct physical intervention** technique that enables nearby pedestrians to stop the AV by tapping on the surface in close-range, low-speed scenarios, unlike prior approaches based on vehicle perception or remote control.
- **Empirical Study of People.** The formative study reveals a **discrepancy** between pedestrians and AV developers regarding external emergency stop mechanisms and derive design requirements grounded in both perspectives.
- **Artifact.** We develop a tapping-based system with a dataset of diverse behaviors and machine learning classifiers integrated into a curated triggering mechanism that distinguishes intentional stop requests from incidental or malicious inputs.
- **Empirical Study of System Use.** We conduct a **field study with a moving vehicle**, simulated emergency scenarios, providing an initial exploration of *Tap2Stop* in a field setting and uncovering user preferences for richer bidirectional feedback.

## 2 Related Work

This section reviews prior work from two perspectives: 1) how AVs communicate information to pedestrians (Sect. 2.1), and 2) how AVs interpret pedestrian intent to anticipate and avoid potential conflicts. For the specific intent of “requesting a stop,” we further discuss existing emergency stop mechanisms (Sect. 2.2).

Existing research primarily focuses on communication from vehicles to pedestrians, where AVs actively convey their state or intention. In contrast, the reverse direction (i.e., pedestrians communicating with AV) is passive. Current approaches rely on sensing systems to observe pedestrian behavior and infer intent (e.g., motion or trajectory prediction), rather than enabling pedestrians to explicitly express their intentions. This gap motivates our work to explore mechanisms that allow pedestrians to directly communicate intent to AVs with an initial focus on stop requests.

## 2.1 Interaction between Automated Vehicles and Pedestrians

Research on AV-pedestrian interaction mainly focused on one-way communication from the AV to pedestrians (with Colley et al. [14] being a notable exception as they allow bi-directional communication), which falls under the domain named external human-machine interfaces (eHMI) [6, 15, 16]. These interfaces are mounted on the AV's exterior surface and communicate with various modalities, such as text displays on windshields [5, 26], blinking light strips around headlights [13, 20], or robotic eyes and arms [9, 27, 28, 40]. The eHMIs typically serve two roles: 1) acting as a "virtual driver" that communicates the AV's status or intention when there is no human driver [41]; and 2) functioning as an "augmented mirror," sharing sensory information beyond human perception ability, such as AVs approaching from blind spots [10, 17, 25].

Although eHMIs improve communication, pedestrians still raise concerns about trust and safety toward both the eHMI and AVs [29, 32]. Post-interviews of past research show two reasons that can explain the concerns: 1) There is no guarantee that the eHMI will always convey accurate information [30, 35], and 2) messages are generated from the AV's perspective and may not always be appropriate for pedestrians (i.e., no social interaction void exists) [19]. That is, one-way communication is insufficient, and pedestrians also need a channel to express their intentions to AVs. The ability to request an emergency stop represents a fundamental need that could significantly improve perceived safety. However, there is a lack of research exploring how pedestrians can convey intent to AVs [7]. In this study, we propose a novel approach that enables pedestrians to communicate with AVs by tapping on them.

## 2.2 Emergency Stop Method For Vehicles

Current emergency stop mechanisms are typically embedded as extended features. Typical implementations are: 1) a physical red emergency stop button located near the passenger seat [36], 2) internal risk identification systems, which extend from the AV's sensing capabilities, from visual to multimodal perception [31, 38, 54], such as sound detection to identify risky events and trigger emergency stops [1], and 3) physical mitigation equipment such as external bumpers in the form of cushions designed to absorb impact as a last resort [34].

However, these systems are **vehicle-centered**: they are highly dependent on the AV's decision-making systems (e.g., route planning). If these fail to recognize the danger or the passenger fails to detect the emergency, the emergency stop systems fail accordingly. To address this limitation, researchers have explored external

emergency stop approaches integrating with wireless communication network technology [55]. These systems collect road traffic information and can stop specific AVs either automatically [21] or through intervention by an authorized monitor (e.g., traffic control centers) [18, 47].

Neither vehicle-centered nor remote-controlled systems allow human pedestrians to directly trigger an emergency stop. That is, in emergency situations, pedestrians currently lack any direct method to communicate with the AV. In this study, we focus on close-range, low-speed scenarios and employ a touch-based interaction modality that enables pedestrians to communicate their intent to the AV. We explore tapping patterns as a direct and intuitive signal for emergency stop requests.

## 3 The Design Criteria

We conducted a formative study with both pedestrians and AV developers. The findings shows pedestrians' needs of pedestrian-activated emergency stop method (Sect. 3.1) and developers' concerns (Sect. 3.2). We then derived "tap" as an explorable interaction modality, together with its key design prerequisites (Sect. 3.3).

### 3.1 Questionnaire with 23 pedestrians

We conducted a questionnaire with pedestrians (N=23), recruited via a public website (<https://sciencepark.jp/>), to better reflect general public perceptions of AVs. None of them had experience in AV development. Participants rated **familiarity with AVs**, **trust in AVs** and the **need for external emergency stop mechanisms** on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree), followed by an open-ended question on **how they would stop a car** in a risky situation.

Results (Figure 2) show moderate trust in AVs ( $M=3.48$ ,  $SD=0.97$ ), while all participants agreed that pedestrians should be able to stop AVs externally ( $M=4.87$ ,  $SD=0.34$ ). Open-ended responses were coded into three sensory categories: *visual* (6 out of 23, e.g., gestures or gaze), *auditory* (10 out of 23, e.g., screaming, shouting or special voice commands), and *taptic* (6 out of 23, e.g., physical contact such as knocking or kicking). Other than human sense, *remote control* methods were suggested (9 out of 23), including smartphone-based control (6 out of 23), where users input a predefined code to stop a nearby AV, and roadside buttons (3 out of 23), inspired by pedestrian crossing signals to stop vehicles at intersections.

### 3.2 Interview with 8 AV developers

We then interviewed 8 AV development researchers, including four senior researchers from academia and industry, and four PhD students with hands-on experience. Participants covered diverse areas, including perception, planning, decision-making, infrastructure cooperation, and simulation. They were asked about **Q1) their trust in AVs** and **Q2) whether pedestrians should be able to influence AV decisions**. (Actual questions are in Appendix A and details description of their status in academia or industry and research field are in Appendix B)

For the **Trust (Q1)**, different from general public who gives over borderline score (3.48/5), developers shows negative tendency (e.g., "Not very much", "As a developer of autonomous vehicle technology, I still don't fully trust self-driving cars due to their

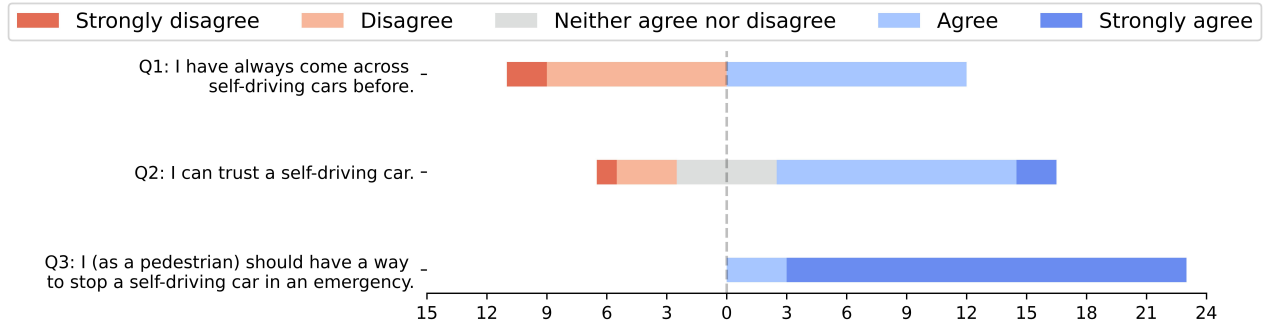


**Figure 2: The Likert chart presents pedestrians’ perspectives on autonomous driving, measuring their familiarity with self-driving cars, trust in the technology, and agreement about the need for pedestrian-initiated stopping capability.**

shortcomings”, and “I trust AV assistants, but I’m not so confident in L4/L5 autonomous driving.”) We found their lower trust due to concerns about sensing and hardware constraints, real-world edge cases beyond operational design domains, limited interaction with vulnerable road users [33], and cybersecurity risks.

For the **Measure of empowering pedestrians (Q2)**, responses fell into two categories: the first was absolutely no (3 of 8), and the second was yes, but with well-defined boundaries (5 of 8). A shared concern was that unbounded control could enable misuse (e.g., abuse or malicious use) and oversimplify interaction into a binary “stop,” failing to capture nuanced intent and potentially leading to conservative or inefficient AV behavior.

In summary, developers acknowledge current AV limitations and the necessity of pedestrian influence for safety, but emphasize that such interaction must prevent misuse and ensure effective communication.

### 3.3 Rationale for choosing “Tap”

To identify a suitable approach for close-range, low-speed scenarios, we progressively narrowed down candidate methods based on participants’ responses.

**From the pedestrian perspective, we applied the design criteria:** 1) accessibility to the general public, excluding system-controlled approaches (e.g., traffic officer control); 2) reliable and fast detectability by AVs, favoring auditory over vision-dependent signals for real-time processing [24]; and 3) intuitive use without prior training, excluding complex or time-consuming actions under emergency conditions. This process identified **tapping** and **shouting** as feasible modalities.

**From the developer perspective, we identified two key concerns:** *system abuse* and *unclear intentions*. To address abuse, the mechanism should resist malicious or unintended activation (e.g., deliberate tapping or environmental noise). To address unclear intentions, it should offer robust features to support intent classification across varying urgency levels. Both **tapping** and **shouting** are distinguishable from environmental noise and provide rich features for classifier training.

We finally selected **tapping** as the emergency stop signal due to its target specificity. That is, a problem with shouting is that target (which vehicle should stop) is not clear when multiple vehicles

are running, while physical contact unambiguously identifies the target vehicle.

## 4 Tap2Stop Development

To put the “Tapping” concept into a practical system, we first set up the **sensing hardware**. We explored multiple approaches for the hardware. Our initial design used an Arduino with multiple sensors (e.g., sound, IMU), but its sampling rate (1 kHz) and resolution (10-bit) were insufficient to capture high-frequency features. We therefore attached a wireless microphone (Ulanzi, 44.1 kHz, 16-bit) to the interior surface of the vehicle (Figure 3, blue box), enabling reliable collection of high-quality raw audio for subsequent signal processing.

### 4.1 Reliability of Tap2Stop

To enable reliable real-time operation of *Tap2Stop*, we identified two key requirements based on the design criteria from the formative study: (1) *Tap2Stop* should accurately distinguish intentional taps from non-intentional and malicious inputs, and (2) *Tap2Stop* should balance detection accuracy with reaction time.

To satisfy the first requirement, we developed and optimized the tap recognition system. We first collected a tapping dataset, then evaluated seven machine learning algorithms, selected the best-performing model, and optimized its parameters to maximize classification performance (Sect. 4.2).

Meeting the second requirement requires more than accurate tap classification. While the classifiers estimate the intention from each observation, an emergency stop is ultimately a decision that should be made under continuous real-time sensing. Unlike offline inference on complete tapping segments, real-time operation continuously evaluates overlapping sliding windows from an incoming stream. At the onset of a tap, the first few windows inevitably capture only partial observations, causing classifier predictions to fluctuate before sufficient evidence becomes available. Triggering an emergency stop directly from a single classifier output therefore risks either false activations or delayed responses. To address this issue, we introduce a gating and confidence accumulation mechanism (Sect. 4.3) that aggregates evidence across consecutive windows before issuing an emergency stop command.

The resulting *Tap2Stop* system was then integrated into a real autonomous vehicle to evaluate how pedestrians perceived and used *Tap2Stop* in real-world AV interactions.

## 4.2 Classifier Specification

We define two classifiers in *Tap2Stop* that correspond to the identified requirements.

**Classifier 1: Tap Detection.** This classifier distinguishes tapping signals from normal driving sounds, ensuring the system is not falsely activated during regular operation.

**Classifier 2: Detection of Malicious Tapping.** This classifier improves robustness by identifying non-intentional or malicious tapping behaviors, enabling the system to respond only to genuine stop requests.

## 4.3 Trigger Mechanism

The *Tap2Stop* trigger mechanism employs two coordinated classifiers (“Malicious” and “Tap”) within a gating-accumulation architecture. The “Malicious” classifier acts as a gate (pass or discard), while the “Tap” classifier operates within an accumulation process that integrates confidence over time. Continuous audio is processed in 0.1 s steps. Segments classified as malicious are discarded, and the remaining segments are evaluated to detect tap events. Valid tapping signals are accumulated until the confidence exceeds a predefined threshold, triggering a “stop” command; otherwise, the emergency stop system continues to wait. This accumulation mechanism helps filter out unintentional taps and environmental noise (e.g., rain or water droplets).

The resulting gated signal is then temporally integrated using exponential decay over a fixed number of past frames:

$$d(t) = \sum_{k=0}^K h(t-k) \cdot \exp(-\lambda k) \quad (1)$$

where  $h(t)$  denotes the gated confidence of the second classifier at frame  $t$ ,  $K$  is the temporal integration window size, and  $\lambda$  controls the decay rate. A stop signal is triggered when  $d(t)$  exceeds a decision threshold  $\tau_D$ . Each classifier output at frame  $t$  is computed from a time window centered at  $t$ , introducing an inherent look-ahead delay of half the window length. The end-to-end latency from tap onset  $t_{\text{pat}}$  to the decision time is therefore:

$$\Delta t = (t_{\text{cross}} - t_{\text{tap}}) + \frac{T_w}{2} \quad (2)$$

where  $t_{\text{cross}}$  is the frame at which  $d(t)$  first exceeds  $\tau_t$ , and  $T_w$  is the analysis window length.

**Parameter Tuning.** There are several parameters that need to be tuned to balance detection accuracy, robustness to noise, and system latency for real-world operation (Sect. 7.1):

- $\tau_m$ : gating threshold for filtering “malicious” detections
- $\tau_t$ : decision threshold for “triggering”
- $\lambda$ : decay rate controlling temporal sensitivity (short-term vs. long-term accumulation)
- $k$ : time-lag index representing the number of frames in the past

To achieve this, we first optimize these parameters in Study 1 through classifier training, and then use the optimized settings in Study 2 for real-world evaluation.

## 5 Study 1: Data Collection for Training

To train the classifier, we collected a tapping dataset by asking participants to tap on a physical vehicle under various conditions.

### 5.1 Dataset Preparation

Training a reliable classifier requires a large and balanced dataset across different conditions. However, collecting equivalent data on an actual moving vehicle would require over hundreds of participants, and the noise conditions (e.g., environmental and driving sounds) cannot be evenly controlled in real-world settings.

Therefore, we adopt a **feasible noise-overlay augmentation method**. We collected tapping data on a stationary vehicle and augmenting it with seven types of noise. This approach allows us to efficiently generate sufficient and balanced samples while maintaining control over noise conditions. Then, the dataset preparation consists of two parts.

For the first part (noise dataset), we pre-recorded seven driving sound, including **seven elements**: (1) normal driving sound, (2) driving backward with a warning alarm, (3) sudden braking, (4) starting to drive, (5) driving with passengers talking inside, (6) passing pedestrians talking outside, and (7) passengers clapping inside the vehicle. These noises were intended to provide balanced training data rather than represent the full range of real-world driving environments.

For the second part, we recruited 23 participants online (the same group as in the formative study questionnaire) and asked them to imagine four scenarios (S1-S4 in Figure 3) and sequentially tap the vehicle surface for each. The four scenarios were designed based on concerns raised by AV developers.

- **S1) High-emergency:** An AV is approaching a looking away child, and may not be detected. The participant taps to stop it.
- **S2) Low-emergency:** The participant has accidentally left a bag on the road. They tap the AV to prevent potential damage.
- **S3) No intent:** The participant taps the slow-moving AV out of curiosity, without any serious or specific intent.
- **S4) Malicious intent:** The participant discovers that tapping may stop the AV and performs the tapping as a prank, intending to trigger an unnecessary emergency stop.

We used a modified golf cart equipped with a sensing setup described in (Sect. 4.1). We recruited 23 participants, among whom 22 provided valid data. Each participant received a \$5 Amazon gift card as compensation. Participants were instructed to tap a **fixed location** on the vehicle surface (Figure 3), positioned directly above the attached microphone, on a stationary vehicle across four scenarios. To enhance ecological validity while maintaining safety, the vehicle remained stationary, and each scenario was verbally introduced with emotional context. Participants were instructed to tap freely, without constraints on the number of taps or gestures to capture natural behavior.

**The final dataset consists of 616 audio clips** (22 participants  $\times$  4 scenarios  $\times$  7 noise conditions), capturing variations across individuals, scenarios, and environments. We overlaid environmental

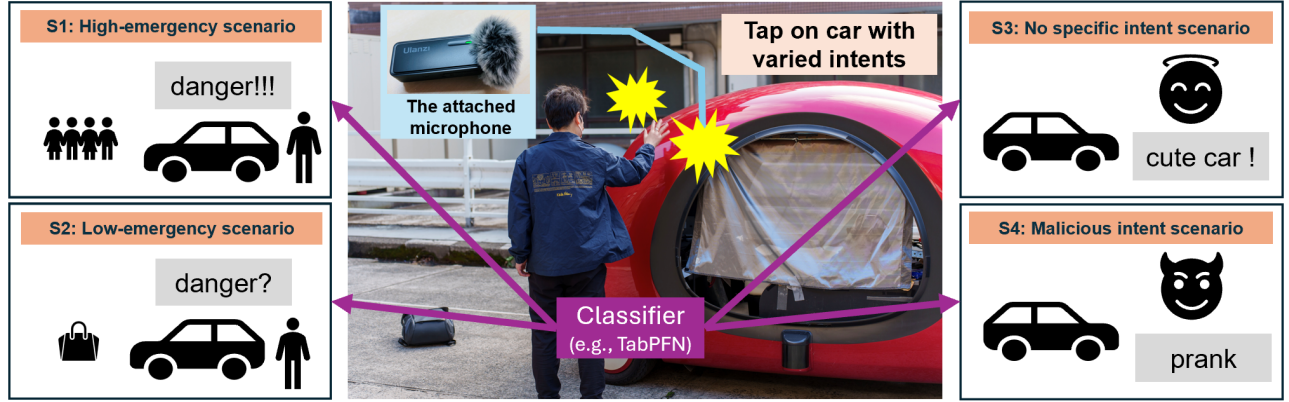


**Figure 3: Data collection setup:** A prototype equipped with an in-vehicle microphone was used to collect data across four conditions (S1: high emergency, S2: low emergency, S3: non-intent, S4: malicious intent).

noise with an average SNR of 18 dB (5th percentile: -4 dB; 95th percentile: 33 dB) to approximate realistic driving conditions.

## 5.2 Classifier Development

We evaluated seven algorithms, ranging from simple to complex. The overlaid tapping audio was categorized into four groups: tapping with 1) strong emergency intent, 2) weak emergency intent, 3) no intent, and 4) malicious intent.

We divided the analysis into four tasks. **Two basic task (1&2) evaluate the technical feasibility** of the two classifiers in the *Tap2Stop* system. **Two advanced tasks (3&4) are progressive classification** that examine whether the system can encode more complex information, such as different types of unintentional tapping or levels of emergency.

- **Basic Tasks:**
  - **Task 1:** Activation stability (Groups 1&2 vs. noise dataset)
  - **Task 2:** Malicious tap detection (Groups 1&2 vs. Groups 3&4)
- **Advanced Tasks:**
  - **Task 3:** Stop-intent detection (Groups 1&2 vs. Group 3)
  - **Task 4:** Emergency intensity classification (Group 1 vs. Group 2)

We first conducted feature analysis in both time and frequency domains [23]. In the time domain [22, 37], we extracted temporal, rhythm, and energy features to capture tap structure and intensity. In the frequency domain [3, 45], we used frequency band and spectral features to characterize energy distribution and temporal variation [50]. 29 sub-features were selected (see Appendix C). We then performed feature augmentation (i.e., polynomial transformations and interactions) and applied Fisher score-based selection [44] to extract the top features for each task (Table 2 in Appendix). The results show that different feature types dominate across tasks: temporal and rhythmic features for malicious detection, and energy/spectral features for emergency-level differentiation. These observations guided the classifier design. We then tested **seven representative models**, ranging from simple (Logistic Regression, Naive Bayes), moderate (SVM, Random Forest), to advanced (LightGBM, CatBoost, TabPFN), using **3-fold stratified cross-validation**.

**Table 1: TabPFN results for the four tasks.** Each column reports the performance of TabPFN on one task.

|                  | Task 1<br>activation<br>stability | Task 2<br>malicious<br>tapping<br>detection | Task 3<br>stop-intent<br>detection | Task 4<br>emergency<br>intensity<br>classification |
|------------------|-----------------------------------|---------------------------------------------|------------------------------------|----------------------------------------------------|
| <b>Accuracy</b>  | <b>1.0000</b><br>(±0.0000)        | <b>0.8117</b><br>(±0.0136)                  | <b>0.7891</b><br>(±0.0400)         | <b>0.8117</b><br>(±0.0199)                         |
| <b>Precision</b> | <b>1.0000</b><br>(±0.0000)        | <b>0.8101</b><br>(±0.0178)                  | <b>0.7760</b><br>(±0.0236)         | <b>0.8121</b><br>(±0.0550)                         |
| <b>Recall</b>    | <b>1.0000</b><br>(±0.0000)        | <b>0.8150</b><br>(±0.0152)                  | <b>0.8123</b><br>(±0.0781)         | <b>0.8242</b><br>(±0.0586)                         |
| <b>F1 Score</b>  | <b>1.0000</b><br>(±0.0000)        | <b>0.8124</b><br>(±0.0124)                  | <b>0.7923</b><br>(±0.0460)         | <b>0.8143</b><br>(±0.0102)                         |

**The accuracy, precision, and recall report** of seven algorithms across four tasks (in Table 3 in Appendix). TabPFN achieved the best performance (in Table 1), with 100% accuracy in distinguishing tapping from driving noise (Task 1) and around 81.17% accuracy in detecting malicious tapping (Task 2), validating system feasibility. For more challenging scenarios (Tasks 3 and 4), such as different types of non-intentional tapping (78.91% accuracy) and varying emergency intensity (81.17% accuracy), performance remain around 80%, indicating that tapping signals encode meaningful contextual information.

## 6 Study 2: Field Study for Validation

We **operated the *Tap2Stop* in physical setting** based on the result in Study 1. According to the accuracy results across 7 classifiers (Table 3 in Appendix), we selected **two independent TabPFN classifiers**. The system processes a continuous audio stream using 1-second windows with a 0.1-second hop.

Due to the window length, it incurs a half-window look-ahead latency. The end-to-end latency is approximately 0.42 s for emergent taps and 0.70 s for warning taps. Based on training, we **select the parameters**  $\tau_m = 0.60$ ,  $\lambda = 0.90$ ,  $\tau_t = 0.88$  and  $K = 20$ , that give the optimization of system performance for Study 2 evaluation.

We then conducted a **Wizard-of-Oz-based field study** with a moving vehicle for two purposes: 1) to validate the system design

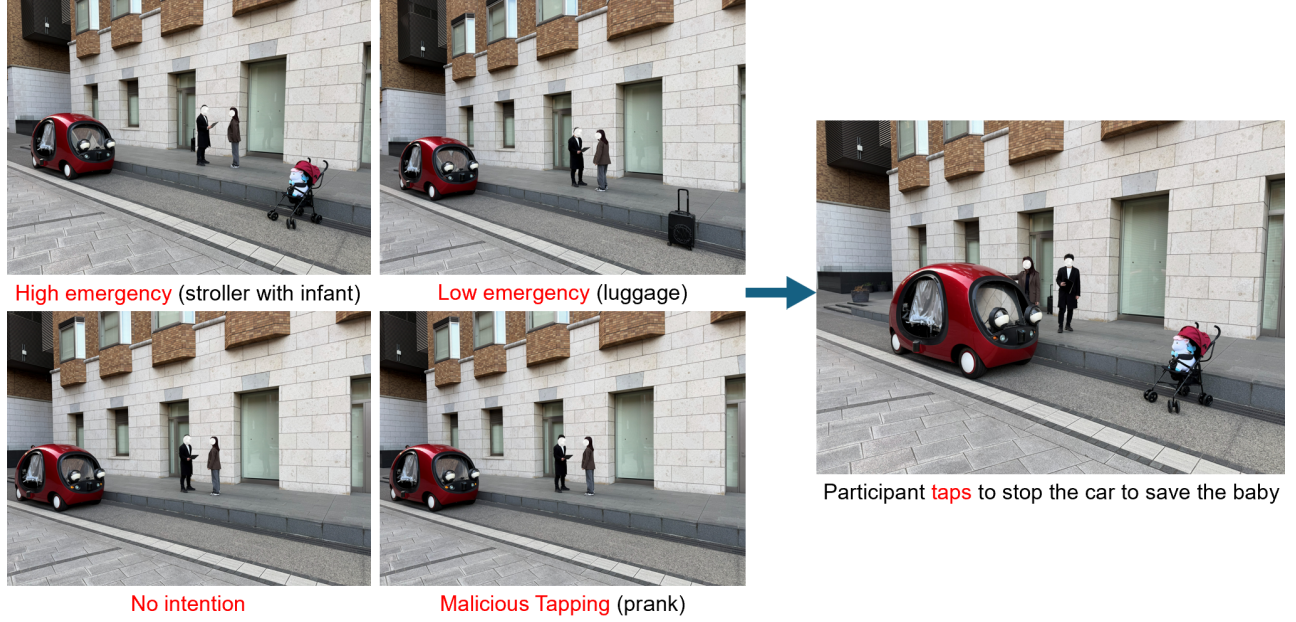


**Figure 4: Field study setup:** The left images show the four scenarios with same four conditions in Study 1 (highlight in red), including the vehicle’s starting point and hazard placement (e.g., an infant stroller). The instructor stood between the participant and the vehicle, partially blocking their view, and faced away as if unaware. As shown on the right, once participants noticed the danger, they interrupted the conversation with instructor and tapped the vehicle to stop it.

and its real-world performance with the above setup; and 2) to capture users’ thoughts and feelings when interacting with a moving vehicle through tapping on a real vehicle surface in a real-world scenario.

## 6.1 Field Study Setup

We recruited a new group of participants ( $N=18$ ) who did not take part in the dataset collection. Each participant received a \$10 Amazon gift card as compensation. To ensure consistency, we used the same vehicle and data collection system, allowing the previously optimized classifier settings to be directly applied. Tapping conditions were also kept consistent (highlighted in red in left part of Figure 4), with a fixed tapping location (Figure 3) while allowing natural variation in gesture, speed, force, and number of taps.

The study was conducted with a **moving vehicle** presented as automated using a Wizard-of-Oz approach. The setup involved three experimenters: a hidden driver (“Wizard”) controlling the vehicle, an instructor interacting with the participant, and a coordinator monitoring data collection and classifier performance.

Participants were instructed that tapping on the vehicle could stop it and were asked to intervene when they perceived a need to do so. To capture natural responses, we **physically recreated realistic scenarios** instead of asking participants to imagine them. We constructed four scenarios (Figure 4, left part) using specific physical objects, following the same definitions as in the dataset: high emergency (vehicle approaching an infant in a stroller), low emergency (approaching a piece of luggage), and non-intent and malicious / prank conditions (no obstacle present).

During each trial, the instructor stood between the participant and the vehicle, partially blocking their view, and faced away from the vehicle while pretending to be unaware of its movement. The participant was unaware of the hidden driver and was told the vehicle was autonomous. After the coordinator confirmed system readiness, the vehicle began moving toward the object (e.g., stroller). The instructor, facing away from the vehicle, continued the conversation with participant without reacting. Once the participant noticed the potential danger, they typically interrupted the conversation, moved past the instructor, and tapped the vehicle to stop it (Figure 4, right part). To ensure safety, two precautions were implemented. First, the vehicle was predefined to stop in all scenarios regardless of participant action or *Tap2Stop* system output. Second, participants stayed on an elevated platform above the wheel’s radius, preventing the vehicle from reaching them.

## 6.2 Measurement

Before the field study, the instructor introduced the concept of *Tap2Stop* (i.e., tapping enables pedestrians to actively stop the AV). Participants then completed a **before-study questionnaire** assessing their general perceptions of autonomous driving and the *Tap2Stop* method using a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree):

- Q1. I feel autonomous driving is safe.
- Q2. I trust autonomous vehicles.
- Q3. I feel autonomous vehicles are alive or have intelligence.
- Q4. I think tapping the car as a stopping method is necessary.

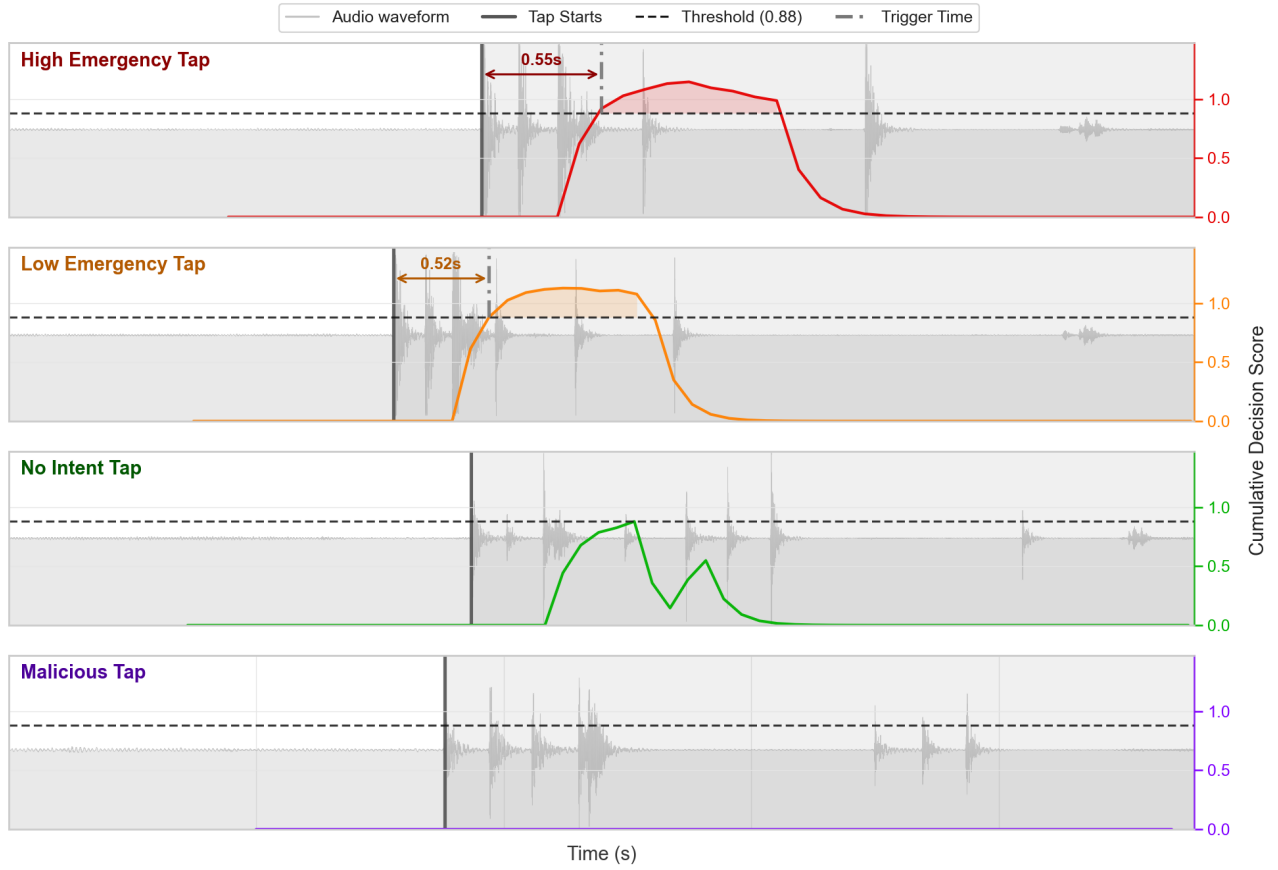


**Figure 5: Decision score illustrating the real-time operation of *Tap2Stop* across four scenarios. The x-axis represents time (s), and the y-axis shows the cumulative decision score. The vertical line indicates tap onset, the dashed line denotes the trigger threshold, and the colored curve shows the accumulated score.**

After completing the four tapping tasks, participants answered the “same” four questions again as **post-study questionnaire**. Then, the participants were asked two open-ended questions:

- Q5. Was the tapping action intuitive to perform?
- Q6. Would you be willing to tap to stop the vehicle in the future?

## 7 Analysis

We analyze the collected data to evaluate system performance and understand users’ behaviors and experiences during real-world interaction. We first report quantitative results to demonstrate reliable performance (Sect. 7.1), followed by an illustrative example showing how the system operates in practice (Sect. 7.2). We then analyze users’ feelings (Sect. 7.3) and thoughts (Sect. 7.4).

### 7.1 Field Study Performance

We analyze the results in two types: 1) intended taps in emergency cases, and 2) unintended or malicious taps in non-emergency cases.

**For Type 1 (intended taps)**, participants attempted to stop the vehicle in emergency situations. In this case, the system is expected to trigger a stop. We therefore report the detection rate. The system

achieved a 100% detection rate for high-emergency taps (object: stroller; latency: 0.42,s) and 71% for low-emergency taps (object: luggage; latency: 0.70,s). The detection rate for high-emergency taps is 100%, suggesting that the system can recognize stop requests under the conditions of our field study (i.e., with a moving vehicle and environmental noise). In contrast, for the low-emergency condition, accuracy is 71%, suggesting that when participants exhibit more nuanced or uncertain intent (e.g., hesitation about whether to stop the vehicle), the resulting tapping patterns are less distinctive, making them more difficult to interpret as a clear stop request.

**For Type 2 (unintended or malicious taps)**, there is no actual emergency, and participants either tapped without intent or with malicious intent. In these cases, the system should not trigger a stop. We therefore report the false positive rate, defined as cases where the system incorrectly triggers a stop. The false positive rate was 29% for no-intent taps and 24% for malicious taps. These rates suggest that while the safeguard against unintended activations is acceptable, it cannot fully eliminate them.

We clarify that the results of the four tasks in Study 2 and Study 1 are not directly comparable, as one reflects a system-level evaluation while the other evaluates a standalone classifier. The results of Study

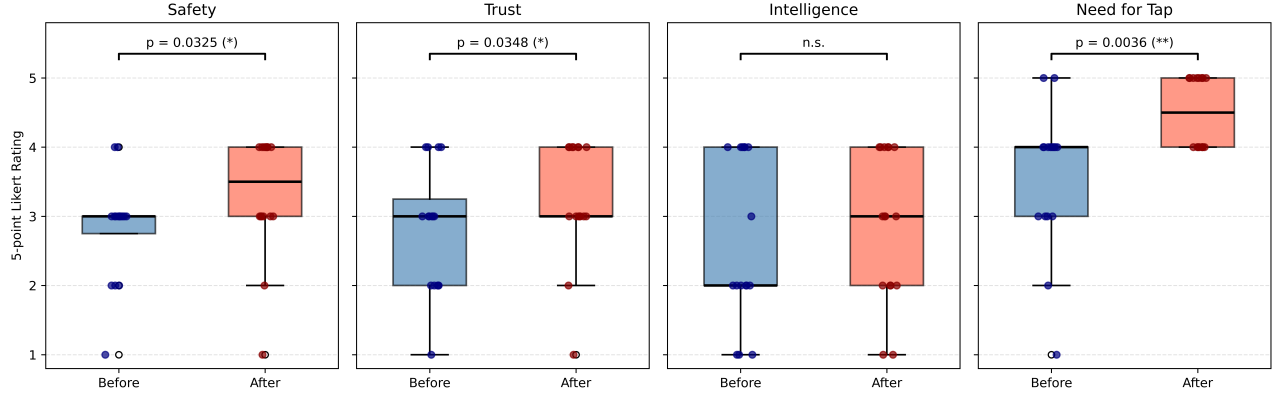


Figure 6: Comparison of participants' ratings before and after the study across four same statements.

2 indicate that the system meets the design requirements. However, the variance of participants, environmental noise, and real-world processing conditions may increase unintended activations.

## 7.2 Case Study: Example of System Operation

We use one participant's (P5's) four tapping data as an example (Figure 5) to illustrate how the *Tap2Stop* gating-accumulation system (as described in Sect. 4.3) operates in real-world setting.

Figure 5 illustrates the temporal evolution of the decision score under four scenarios. The gray waveform indicates the input audio signal. The vertical black line marks the tap onset, the dashed line indicates the trigger threshold, and the colored curve shows the accumulated score over time. For high-emergency (red curve) and low-emergency (orange curve) taps, the accumulated score exceeds the threshold, triggering the alert. In the no-intent case (green curve), although some taps are detected and partially accumulated, the score remains below the threshold, and no trigger is issued. In the malicious case (purple curve), the malicious classifier filters inputs before accumulation, resulting in no score increase and no trigger. These results show that the system distinguishes valid stop requests from non-intentional or malicious inputs while remaining robust to false activations.

## 7.3 Likert Rating before vs after "Tapping"

The questionnaire analyses compared participants' ratings before and after interacting with the *Tap2Stop* system across four aspects (perceived safety, trust in AVs, perceived intelligence, and the necessity of tapping as a stopping method) by asking the same four questions twice.

We first assessed the normality of the paired differences using the Shapiro-Wilk test. As the normality assumption was violated, non-parametric Wilcoxon signed-rank tests were used for all questionnaire items.

Overall, ratings increased after the study for all four items (Figure 6). Statistically significant improvements were observed in participants' perceived safety when interacting with the AV ( $p = .0325$ ), trust in the AV ( $p = .0348$ ), and perceived necessity of tapping ( $p = .0036$ ), indicating that direct interaction with the system positively influenced participants' attitudes.

That is, although participants were verbally informed that tapping could stop the vehicle, directly experiencing the interaction (i.e., tapping the car and observing it stop) provided a stronger sense of safety and trust, increasing their preference for such a stopping method. In contrast, the increase in perceived intelligence was not statistically significant ( $p = .2795$ ).

## 7.4 Thematic Analysis on Thoughts

One experimenter first conducted an inductive analysis of the responses to the two open-ended questions. Responses were first coded using **keywords and short phrases** capturing participants' main reasons (e.g., instinctive, human-driver habit, effective, afraid of damaging). **Related codes** were grouped into candidate themes explaining participants' reasons. For reporting purposes, the **final themes** were organized according to whether they reflected supportive or non-supportive views toward each question (i.e., reasons supporting or opposing the intuitiveness of tapping and the willingness to use). The coding results, candidate themes, and theme names were then reviewed by the other two experimenters, and disagreements were resolved through discussion until consensus was reached. **Representative quotes** were selected to illustrate each final theme, as listed below.

The first is the **Intuitiveness of Tapping**. Most participants (13/16) considered tapping an intuitive way to request a stop. This perception was largely attributed to the transfer of behaviors from human-driver interactions, where similar actions are used to attract attention (e.g., "tapping or shouting to stop a car is instinctive, naturally transferred to an unmanned scenario" (P2)). In addition, several participants described tapping as a hand-first reaction under urgency, requiring little deliberate thought (e.g., "When trying to stop it, the hand comes out first" (P8)). However, a minority of participants (3/16) reported that tapping was not intuitive, as it conflicted with their personal instincts, such as pushing or shouting instead. These results suggest that while tapping aligns with common human behaviors and reflexes, individual differences in instinctive reactions remain.

The second is the **Willingness to Use Tapping** in the future. Most participants (12/16) expressed willingness to use tapping as a stopping method, describing it as immediate, simple, and effective

(e.g., “The fastest and most effective behavior is to directly tap the car” (P3)). Beyond its practicality, participants also valued it as a last-resort safety option that provides psychological reassurance in emergencies (e.g., “it leaves me a last chance to protect myself, which gives me peace of mind” (P3)). However, a minority of participants (4/16) were hesitant due to concerns about potential damage to the vehicle and their own safety (e.g., “I am afraid to break the car surface” (P5)). Some also noted that if the tap failed to stop the vehicle, being close enough to interact could increase risk. These findings suggest that while tapping is broadly acceptable as an emergency interaction, perceived risks and system reliability remain critical factors influencing adoption.

## 8 Discussion

We first examine the performance of the system under various scenarios (Sect. 8.1). We then discuss two future directions from participants’ responses: 1) the role of tap location in interaction design (Sect. 8.2) and 2) the need for bidirectional communication (Sect. 8.3), where AVs provide feedback to pedestrians after their tapping.

### 8.1 Performance Across All Tasks

We found an interesting pattern that the detection of malicious tapping outperformed non-intentional tapping. We speculate on the reasons from two perspectives.

**From an interaction perspective,** we think of two possible explanations. First, intentional actions (e.g., malicious or deliberate taps) may produce more structured and distinctive patterns, potentially due to increased cognitive load [51], making them easier to detect. Second, we observed that some participants exaggerated their taps (e.g., more forceful), creating patterns that differ from true emergency tapping. In contrast, unintentional taps tend to be more random and less structured [52], often resembling low-emergency cases, and are therefore harder to distinguish.

**From a machine learning perspective,** this is supported by the distribution overlap analysis. Malicious tapping shows a lower Bhattacharyya coefficient (0.295) and a higher Hellinger distance (0.834) compared to non-intentional tapping (0.503 and 0.648, respectively), indicating less overlap and greater separability from intentional patterns. In contrast, the higher overlap of non-intentional tapping makes it more difficult to classify.

In addition, as for the model complexity, the results indicate that more expressive models (e.g., TabPFN) better capture differences in emergency intensity from tapping signals. To support more nuanced intent recognition, future systems may require more powerful models, such as deep neural networks.

### 8.2 Tapping Location and Localization

In both Study 1 and Study 2, participants were instructed to tap a fixed location on the vehicle surface, directly above the attached microphone. This setup led participants to express the need for greater flexibility in tapping location. Participants reported increased trust and willingness to engage, but also highlighted the importance of **whole-vehicle detectability and spatial localization**.

Participants noted that, in emergencies, they would instinctively tap any reachable part of the vehicle rather than a specific location

(e.g., “I will instinctively tap whatever place I can reach.” (P11), “In an emergency it’s hard to control which exact part of the car you hit.” (P2)), suggesting the need for full-surface detectability. They also expected the vehicle to infer tap location [11], as it may indicate the source of danger. While our current prototype uses a single microphone, future systems could incorporate multi-microphone arrays across the entire vehicle surface to improve accuracy and enable localization of tapping sources.

### 8.3 Bi-directional Communication

During the open-ended responses, participants expressed a need for feedback beyond vehicle stopping. We grouped these into three categories with representative quotations. **1) Safety-related feedback.** Many participants (11/16) requested confirmation that their tapping was detected to avoid repeated actions, and seven wanted the vehicle to indicate its understanding (e.g., “I detected a person ahead”). **2) Future action feedback.** A majority (12/16) preferred information about the vehicle’s next action (e.g., “I will stop for 10 seconds”, “Please tap again when it is safe to restart”). **3) Accountability concerns.** Participants also raised responsibility concerns (e.g., “Will I be blamed?”, “Do I need to compensate?”), suggesting the need for reassurance (e.g., “No responsibility required.”).

**We see an opportunity to connect Tap2Stop to the emerging field of eHMI design [6],** including light-based signals (e.g., headlights, LEDs), text displays, audio feedback, motion cues, and haptic feedback [12, 43, 48]. For example, one participant suggested a haptic response, similar to a smartphone home button, to confirm that the vehicle has received their intent (P7). Such bidirectional communication can reduce information asymmetry and enhance pedestrians’ safety and trust in AVs.

## 9 Conclusion

We propose an intuitive, human-centered concept that enables pedestrians to actively stop AVs by tapping on the vehicle surface. We translate this “Tap” concept into a practical system, *Tap2Stop*, through a formative study that identifies key design requirements.

We then develop and evaluate the system through two studies. Study 1 validates technical feasibility by collecting tapping data and training classifiers. Study 2 investigates the usability through a field study with a physical prototype and real participants.

Our results further suggest that tapping can encode subtle differences in emergency intensity, highlighting its potential as an expressive interaction channel. We envision future work exploring *Tap2Stop* toward enabling safer and richer communication channels between autonomous vehicles and pedestrians.

### Open Science

Source code is available at: <https://github.com/ApiXia/tap2stop>

### Acknowledgments

This work was supported by JST CRONOS, Grant Number JP-MJCS24K8, Japan. Mark Colley was supported by a Canon Research Fellowship. We thank all study participants. We thank Ochanomizu University for supporting Study 1 by providing the data collection venue and participant recruitment site.

## References

- [1] Usman Afzal and Lubna Farhi. 2024. Automatic Identification of Emergency Sounds for Self-Driving Cars. *Pakistan Journal of Engineering and Technology* 7, 01 (2024), 22–33.
- [2] Aljazeera. 2026. *US opens probe after a Waymo self-driving car hit a child near a school*. Retrieved Mar 22, 2026 from <https://www.aljazeera.com/economy/2026/1/29/us-opens-probe-after-a-waymo-self-driving-car-hit-a-child-near-a-school>
- [3] Xu Dong An and Zhou Ruan. 2021. Speech Emotion Recognition algorithm based on deep learning algorithm fusion of temporal and spatial features. In *Journal of Physics: Conference Series*, Vol. 1861. IOP Publishing, 012064.
- [4] NBC BAY AREA. 2024. *Cruise admits to filing false report after robotaxi dragged a San Francisco pedestrian*. Retrieved Mar 18, 2025 from <https://www.nbcbayarea.com/news/local/san-francisco/cruise-admits-filing-false-report-robotaxi-dragged-san-francisco-pedestrian/3710410/>
- [5] Pavlo Bazilinskyy, Lars Kooijman, Dimitra Dodou, and Joost CF de Winter. 2021. How should external Human-Machine Interfaces behave? Examining the effects of colour, position, message, activation distance, vehicle yielding, and visual distraction among 1,434 participants. *Applied ergonomics* 95 (2021), 103450.
- [6] Juan Carmona, Carlos Guindel, Fernando Garcia, and Arturo De La Escalera. 2021. eHMI: Review and guidelines for deployment on autonomous vehicles. *Sensors* 21, 9 (2021), 2912.
- [7] Melissa Cefkin, Jingyi Zhang, Erik Stayton, and Erik Vinkhuyzen. 2019. Multi-methods research to examine external HMI for highly automated vehicles. In *HCI in Mobility, Transport, and Automotive Systems: First International Conference, MobiTAS 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings 21*. Springer, 46–64.
- [8] cepton. 2024. *Reading Lidar Specs (Part I): What They Do and Don't Tell You*. Retrieved Aug 7, 2025 from <https://www.cepton.com/driving-lidar/reading-lidar-specs-part-i-what-they-do-and-dont-tell-you>
- [9] Chia-Ming Chang, Koki Toda, Xinyue Gui, Stela H Seo, and Takeo Igarashi. 2022. Can eyes on a car reduce traffic accidents?. In *Proceedings of the 14th international conference on automotive user interfaces and interactive vehicular applications*. 349–359.
- [10] Xu Chen, Xiaomeng Li, Yuxuan Hou, Wenzhang Yang, Changyin Dong, and Hao Wang. 2025. Effect of eHMI-equipped automated vehicles on pedestrian crossing behavior and safety: A focus on blind spot scenarios. *Accident Analysis & Prevention* 212 (2025), 107915.
- [11] Hyunung Cho, Alexander Wang, Divya Kartik, Emily Liying Xie, Yukang Yan, and David Lindlbauer. 2024. Aupimize: Optimal placement of spatial audio cues for extended reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–14.
- [12] Ashley Colley, Jonna Häkklä, Meri-Tuulia Forsman, Bastian Pflöging, and Florian Alt. 2018. Car exterior surface displays: Exploration in a real-world context. In *Proceedings of the 7th acm international symposium on pervasive displays*. 1–8.
- [13] Mark Colley, Elvedin Bajrovic, and Enrico Rukzio. 2022. Effects of Pedestrian Behavior, Time Pressure, and Repeated Exposure on Crossing Decisions in Front of Automated Vehicles Equipped with External Communication. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 367, 11 pages. doi:10.1145/3491102.3517571
- [14] Mark Colley, Jan Henry Belz, and Enrico Rukzio. 2021. Investigating the Effects of Feedback Communication of Autonomous Vehicles. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Leeds, United Kingdom) (*AutomotiveUI '21*). Association for Computing Machinery, New York, NY, USA, 263–273. doi:10.1145/3409118.3475133
- [15] Mark Colley, Julian Britten, and Enrico Rukzio. 2023. Scalability in External Communication of Automated Vehicles: Evaluation and Recommendations. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2, Article 51 (June 2023), 26 pages. doi:10.1145/3596248
- [16] Mark Colley, Pascal Jansen, Mugdha Keskar, and Enrico Rukzio. 2025. Improving External Communication of Automated Vehicles Using Bayesian Optimization. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 80, 16 pages. doi:10.1145/3706598.3714187
- [17] Mark Colley, Surong Li, and Enrico Rukzio. 2021. Increasing Pedestrian Safety Using External Communication of Autonomous Vehicles for Signalling Hazards. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction* (Toulouse & Virtual, France) (*MobileHCI '21*). Association for Computing Machinery, New York, NY, USA, Article 20, 10 pages. doi:10.1145/3447526.3472024
- [18] Mark Colley, Jonathan Westhauser, Jonas Andersson, Alexander G. Mirnig, and Enrico Rukzio. 2025. Introducing ROADS: A Systematic Comparison of Remote Control Interaction Concepts for Automated Vehicles at Road Works. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 81, 20 pages. doi:10.1145/3706598.3713476
- [19] Joost de Winter and Dimitra Dodou. 2022. External human-machine interfaces: Gimmick or necessity? *Transportation research interdisciplinary perspectives* 15 (2022), 100643.
- [20] Debargha Dey, Azra Habibovic, Bastian Pflöging, Marieke Martens, and Jacques Terken. 2020. Color and animation preferences for a light band eHMI in interactions between automated vehicles and pedestrians. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
- [21] Fabian Duerr, Jens Ziehn, Ralf Kohlhaas, Masoud Roschani, Miriam Ruf, and Jürgen Beyerer. 2020. Realtime global optimization of a fail-safe emergency stop maneuver for arbitrary electrical/electronic failures in automated driving. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 1–8.
- [22] Daniel PW Ellis. 2007. Beat tracking by dynamic programming. *Journal of New Music Research* 36, 1 (2007), 51–60.
- [23] Behnaz Ghoraani and Sridhar Krishnan. 2011. Time-frequency matrix feature extraction and classification of environmental audio signals. *IEEE transactions on audio, speech, and language processing* 19, 7 (2011), 2197–2209.
- [24] Jun Gong, Aakar Gupta, and Hrvoje Benko. 2020. Acustico: Surface tap detection and localization using wrist-based acoustic tdoa sensing. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 406–419.
- [25] Xinyue Gui, Chia-Ming Chang, Stela H Seo, Koki Toda, and Takeo Igarashi. 2024. Scenarios Exploration: How AR-Based Speech Balloons Enhance Car-to-Pedestrian Interaction. In *International Conference on Human-Computer Interaction*. Springer, 223–230.
- [26] Xinyue Gui, Ehsan Javanmardi, Stela Hanbyeol Seo, Vishal Chauhan, Chia-Ming Chang, Manabu Tsukada, and Takeo Igarashi. 2024. "Text+ Eye" on Autonomous Taxi to Provide Geospatial Instructions to Passenger. In *Proceedings of the 12th International Conference on Human-Agent Interaction*. 429–431.
- [27] Xinyue Gui, Mikiya Kusunoki, Bofei Huang, Stela Hanbyeol Seo, Chia-Ming Chang, Haoran Xie, Manabu Tsukada, and Takeo Igarashi. 2024. Shrinkable arm-based ehmi on autonomous delivery vehicle for effective communication with other road users. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 305–316.
- [28] Xinyue Gui, Koki Toda, Stela Hanbyeol Seo, Chia-Ming Chang, and Takeo Igarashi. 2022. "I am going this way": Gazing Eyes on Self-Driving Car Show Multiple Driving Directions. In *Proceedings of the 14th international conference on automotive user interfaces and interactive vehicular applications*. 319–329.
- [29] Xinyue Gui, Koki Toda, Stela Hanbyeol Seo, Felix Martin Eckert, Chia-Ming Chang, Xiang'Anthony Chen, and Takeo Igarashi. 2023. A field study on pedestrians' thoughts toward a car with gazing eyes. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.
- [30] Xinyue Gui, Ding Xia, Mark Colley, Stela Hanbyeol Seo, Chia-Ming Chang, Ehsan Javanmardi, Manabu Tsukada, and Takeo Igarashi. 2026. Exploring LLMs for Generating Communicational Actions of External Interfaces on Autonomous Vehicles. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 10, 2 (2026), 1–28.
- [31] Umar Zakir Abdul Hamid, Fakhru Razi Ahmad Zakuan, Khairul Akmal Zulkepli, Muhammad Zulfaqr Azmi, Hairi Zamzuri, Mohd Azizi Abdul Rahman, and Muhammad Aizzat Zakaria. 2017. Autonomous emergency braking system with potential field risk assessment for frontal collision mitigation. In *2017 IEEE conference on systems, process and control (icspc)*. IEEE, 71–76.
- [32] Ann-Christin Hensch, Isabel Kreißig, Matthias Beggiano, and Josef F Krems. 2021. The Effects of eHMI Failures on Elderly Participants' Assessment of Automated Vehicle Communication Signals. In *International Conference on Human Interaction and Emerging Technologies*. Springer, 355–363.
- [33] Kai Holländer, Mark Colley, Enrico Rukzio, and Andreas Butz. 2021. A Taxonomy of Vulnerable Road Users for HCI Based On A Systematic Literature Review. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 158, 13 pages. doi:10.1145/3411764.3445480
- [34] Qazi Hamza Jan and Karsten Berns. 2021. Safety-configuration of Autonomous Bus in Pedestrian Zone. In *VEHITS*. 698–705.
- [35] Anees Ahamed Kaleefathullah, Natasha Merat, Yee Mun Lee, Yke Bauke Eisma, Ruth Madigan, Jorge Garcia, and Joost de Winter. 2022. External human-machine interfaces can be misleading: An examination of trust development and misuse in a CAVE-based pedestrian simulation environment. *Human factors* 64, 6 (2022), 1070–1085.
- [36] Mihai Kocsis, Nico Sußmann, Johannes Buyer, and Raoul Zöllner. 2017. Safety concept for autonomous vehicles that operate in pedestrian areas. In *2017 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 841–846.
- [37] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 12, 2 (2011), 74–82.
- [38] Surim Kwon, Changyoung Jung, Taesung Choi, Youngchul Oh, and Byungyong You. 2014. Autonomous emergency stop system. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*. IEEE, 444–449.
- [39] Karin Lagerkvist and Li Xiaoyan. 2024. Design of a Remote-Stop device for autonomous vehicles: Interaction design for human-autonomous-systems.
- [40] Yuan Li, Xinyue Gui, Ding Xia, Mark Colley, and Takeo Igarashi. 2025. TailCue: Exploring Animal-inspired Robotic Tail for Automated Vehicles Interaction. In

*Proceedings of the 13th International Conference on Human-Agent Interaction.* 114–123.

- [41] Karthik Mahadevan, Sowmya Somanath, and Ehud Sharlin. 2018. Communicating awareness and intent in autonomous vehicle-pedestrian interaction. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.
- [42] CBS NEWS. 2024. *Tesla's 'Full Self-Driving' system faces probe after pedestrian death*. Retrieved Mar 18, 2025 from <https://www.cbsnews.com/news/tesla-fsd-self-driving-autopilot-elon-musk/>
- [43] patentlyapple. 2023. *Apple invents an advanced External Display System for Autonomous Passenger Vehicles that communicates with the public*. Retrieved Aug 7, 2025 from <https://www.patentlyapple.com/2023/07/apple-invents-an-advanced-external-display-system-for-autonomous-passenger-vehicles-that-communicates-with-the-public.html>
- [44] Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* 27, 8 (2005), 1226–1238.
- [45] Benjamín Pla, Joaquín De la Morena, Pau Bares, and Alexandra Aramburu. 2023. An unsupervised machine learning technique to identify knock from a knock signal time-frequency analysis. *Measurement* 211 (2023), 112669.
- [46] robosense. 2020. *Three Ways for Self-Driving Car's Near-field Blind Spots Detection*. Retrieved Aug 7, 2025 from <https://www.robosense.ai/en/news-show-1325>
- [47] Young-Jae Ryoo. 2015. Remote Emergency Stop System to Improve Safety of Automated Driving Vehicle. *Journal of Institute of Control, Robotics and Systems* 21, 3 (2015), 194–198.
- [48] Xiyuan Shen, Seokhyun Hwang, Junhan Kong, Alexandre LS Filipowicz, Andrew Best, Jean Costa, Scott Carter, James Fogarty, and Jacob O Wobbrock. 2025. Touchscreens in Motion: Quantifying the Impact of Cognitive Load on Distracted Drivers. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–21.
- [49] Dhinesh Subramanian. 2017. *Research in development of emergency stop assistance*. Ph. D. Dissertation. Kauno technologijos universitetas.
- [50] George Tzanetakis and Perry Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing* 10, 5 (2002), 293–302.
- [51] Aldert Vrij, Ronald Fisher, Samantha Mann, and Sharon Leal. 2006. Detecting deception by manipulating cognitive load. *Trends in cognitive sciences* 10, 4 (2006), 141–142.
- [52] Aldert Vrij, Pär Anders Granhag, Samantha Mann, and Sharon Leal. 2011. Outsmarting the liars: Toward a cognitive lie detection approach. *Current Directions in Psychological Science* 20, 1 (2011), 28–32.
- [53] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research Contributions in Human-Computer Interaction. *Interactions* 23, 3 (apr 2016), 38–44. doi:10.1145/2907069
- [54] Lisheng Yang, Tomonari Furukawa, Lei Zuo, and Zachary Doerzaph. 2020. Control Design for Automated Vehicle Emergency Safe Stop System Based on Differential Dynamic Programming. In *Dynamic Systems and Control Conference*, Vol. 84270. American Society of Mechanical Engineers, V001T02A006.
- [55] Yang Yu and Sanghwan Lee. 2022. Remote driving control with real-time video streaming over wireless networks: Design and evaluation. *IEEE Access* 10 (2022), 64920–64932.

## Appendix

### A Interview questions

- **Q1:** How much do you trust automated vehicles? Why? (E.g., believing that current autonomous driving technology rarely makes mistakes.)
- **Q2:** Do you think pedestrians should be able to change automated vehicle decisions? Why? (For example, pedestrians could influence the automated vehicle's decision-making by standing in front of it, making gestures, or touching it. How would this impact automated vehicle operations?)

### B Demographics of Interviewees

The interviewees were asked to provide a brief description of their background and research interests. No personally identifiable information was collected.

#### B.1 Background

##### Question

Please provide a short description of your background (no personal information needed). It would be great to receive information about your position, the duration of your current position, and your experience level.

##### Responses (8)

- (1) AV developer in academia.
- (2) Autonomous vehicle software engineer (3+ years in industry), focusing on V2X and connected solutions.
- (3) PhD student with experience debugging real-world autonomous vehicles.
- (4) 10 months of industry experience.
- (5) AV R&D manager in industry (5+ years of experience), with over 20 years of academic research in mobile robotics.
- (6) PhD student with a 3-month industry internship.
- (7) Project Assistant Professor.
- (8) PhD student currently undertaking an internship.

### B.2 Research Interests

##### Question

Please describe your research interests (i.e., which aspects of automated driving you are working on).

##### Responses (8)

- (1) End-to-end driving, VLM, perception, planning, decision-making, and network communication.
- (2) Vehicle-to-Infrastructure (V2I) cooperation and Collective Perception Systems (CPS) for autonomous driving.
- (3) Cooperative end-to-end autonomous driving.
- (4) Simulation and LLM4AD.
- (5) Full-stack end-to-end systems with a hybrid safety architecture.
- (6) LLM, AI, and decision-making.
- (7) Autoware simulation.
- (8) Infrastructure.

### C Sub-features selection

- **Temporal Features:** Zero Crossing Rate, Temporal Centroid
- **Energy Features:** RMS Energy, Energy Entropy, Short-Time Energy
- **Rhythm Features:** Tempo, Beat Frames, Rhythm Pattern, Onset Strength
- **Frequency Band Features:** Low Frequency (Energy Mean, Energy Std, Centroid), Mid Low Frequency (Energy Mean, Energy Std, Centroid), Mid High Frequency (Energy Mean, Energy Std, Centroid), High Frequency (Energy Mean, Energy Std, Centroid)
- **Spectral Features:** Spectral Centroid, Spectral Rolloff Frequency, Spectral Bandwidth, Spectral Flatness, Spectral Contrast (Mean & Std), Spectral Kurtosis, Spectral Skewness

**Table 2: Top 5 selected augmented features with Fisher scores for each task**

| Task 1: activation stability (Groups 1&2 vs. Pure driving sound) |                                                 |           | Task 3: stop-intent detection (Groups 1&2 vs. 3) |                           |          |
|------------------------------------------------------------------|-------------------------------------------------|-----------|--------------------------------------------------|---------------------------|----------|
| Rank                                                             | Feature (Score)                                 |           | Rank                                             | Feature (Score)           |          |
| Top 1                                                            | RMS Energy                                      | (10.4874) | Top 1                                            | Mid High Freq. Energy Std | (0.8789) |
| Top 2                                                            | Low Freq. Energy Std × Spectral Skewness        | (5.0862)  | Top 2                                            | High Freq. Centroid       | (0.3954) |
| Top 3                                                            | Spectral Kurtosis                               | (3.3833)  | Top 3                                            | Low Freq. Energy Std      | (0.3549) |
| Top 4                                                            | High Freq. Centroid                             | (3.1990)  | Top 4                                            | Spectral Kurtosis         | (0.3049) |
| Top 5                                                            | Short-Time Energy<br>Mid High Freq. Energy Mean | (3.1863)  | Top 5                                            | Low Freq. Centroid        | (0.3044) |

| Task 2: malicious tapping detection (Groups 1&2 vs. 3&4) |                                            |          | Task 4: sentimental variation classification (Group 1 vs. 2) |                                          |          |
|----------------------------------------------------------|--------------------------------------------|----------|--------------------------------------------------------------|------------------------------------------|----------|
| Rank                                                     | Feature (Score)                            |          | Rank                                                         | Feature (Score)                          |          |
| Top 1                                                    | Mid High Freq. Energy Std                  | (0.4697) | Top 1                                                        | RMS Energy                               | (0.3995) |
| Top 2                                                    | Spectral Kurtosis                          | (0.3236) | Top 2                                                        | High Freq. Centroid × Low Freq. Centroid | (0.3669) |
| Top 3                                                    | Spectral Kurtosis<br>High Freq. Energy Std | (0.3159) | Top 3                                                        | High Freq. Energy Std <sup>2</sup>       | (0.3162) |
| Top 4                                                    | High Freq. Centroid                        | (0.3003) | Top 4                                                        | Mid Low Freq. Energy Std                 | (0.2773) |
| Top 5                                                    | Mid Low Freq. Energy Mean                  | (0.2767) | Top 5                                                        | High Freq. Centroid                      | (0.2568) |

**Table 3: Accuracy, precision, recall, and F1 score for four tasks across seven classifiers. The first two classifier columns represent simple baseline models, the middle two represent moderately complex models, and the last three represent more advanced models. Bold values indicate the best performance for each task.**

|                                                                              | Logistic Regression     | Naive Bayes             | SVM                     | Random Forest           | LightGBM                | CatBoost                | TabPFN                  |
|------------------------------------------------------------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| Task 1: activation stability (Groups 1 & 2 vs. various environmental noises) |                         |                         |                         |                         |                         |                         |                         |
| Accuracy                                                                     | 0.9935 (±0.0046)        | 0.9968 (±0.0046)        | 0.9935 (±0.0046)        | 0.9935 (±0.0046)        | 0.9935 (±0.0092)        | <b>1.0000 (±0.0000)</b> | <b>1.0000 (±0.0000)</b> |
| Precision                                                                    | 0.9873 (±0.0090)        | 0.9936 (±0.0091)        | 0.9873 (±0.0090)        | 0.9873 (±0.0090)        | 0.9877 (±0.0175)        | <b>1.0000 (±0.0000)</b> | <b>1.0000 (±0.0000)</b> |
| Recall                                                                       | <b>1.0000 (±0.0000)</b> | <b>1.0000 (±0.0000)</b> | <b>1.0000 (±0.0000)</b> | <b>1.0000 (±0.0000)</b> | <b>1.0000 (±0.0000)</b> | <b>1.0000 (±0.0000)</b> | <b>1.0000 (±0.0000)</b> |
| F1 Score                                                                     | 0.9936 (±0.0045)        | 0.9968 (±0.0046)        | 0.9936 (±0.0045)        | 0.9936 (±0.0045)        | 0.9937 (±0.0089)        | <b>1.0000 (±0.0000)</b> | <b>1.0000 (±0.0000)</b> |
| Task 2: malicious tapping detection (Groups 1 & 2 vs. Groups 3 & 4)          |                         |                         |                         |                         |                         |                         |                         |
| Accuracy                                                                     | 0.6754 (±0.0143)        | 0.6770 (±0.0257)        | 0.7079 (±0.0280)        | 0.7614 (±0.0341)        | 0.7646 (±0.0160)        | 0.7825 (±0.0179)        | <b>0.8117 (±0.0136)</b> |
| Precision                                                                    | 0.6762 (±0.0069)        | 0.7441 (±0.0434)        | 0.7228 (±0.0154)        | 0.7571 (±0.0330)        | 0.7652 (±0.0275)        | 0.7879 (±0.0073)        | <b>0.8101 (±0.0178)</b> |
| Recall                                                                       | 0.6722 (±0.0359)        | 0.5422 (±0.0292)        | 0.6723 (±0.0596)        | 0.7696 (±0.0362)        | 0.7663 (±0.0203)        | 0.7729 (±0.0350)        | <b>0.8150 (±0.0152)</b> |
| F1 Score                                                                     | 0.6738 (±0.0217)        | 0.6266 (±0.0280)        | 0.6958 (±0.0393)        | 0.7633 (±0.0344)        | 0.7652 (±0.0120)        | 0.7800 (±0.0213)        | <b>0.8124 (±0.0124)</b> |
| Task 3: stop-intent detection (Groups 1 & 2 vs. Group 3)                     |                         |                         |                         |                         |                         |                         |                         |
| Accuracy                                                                     | 0.6784 (±0.0362)        | 0.7173 (±0.0614)        | 0.7045 (±0.0213)        | 0.7013 (±0.0202)        | 0.7079 (±0.0353)        | 0.7340 (±0.0469)        | <b>0.7891 (±0.0400)</b> |
| Precision                                                                    | 0.6706 (±0.0463)        | 0.7194 (±0.0754)        | 0.7216 (±0.0300)        | 0.7050 (±0.0300)        | 0.6891 (±0.0176)        | 0.7315 (±0.0195)        | <b>0.7760 (±0.0236)</b> |
| Recall                                                                       | 0.7146 (±0.0304)        | 0.7333 (±0.0661)        | 0.6691 (±0.0291)        | 0.6947 (±0.0200)        | 0.7534 (±0.0816)        | 0.7339 (±0.1132)        | <b>0.8123 (±0.0781)</b> |
| F1 Score                                                                     | 0.6903 (±0.0231)        | 0.7224 (±0.0510)        | 0.6936 (±0.0196)        | 0.6994 (±0.0185)        | 0.7186 (±0.0461)        | 0.7296 (±0.0646)        | <b>0.7923 (±0.0460)</b> |
| Task 4: sentimental variation classification (Group 1 vs. Group 2)           |                         |                         |                         |                         |                         |                         |                         |
| Accuracy                                                                     | 0.6397 (±0.0225)        | 0.6560 (±0.0438)        | 0.6722 (±0.0187)        | 0.7827 (±0.0503)        | 0.7403 (±0.0157)        | 0.7696 (±0.0350)        | <b>0.8117 (±0.0199)</b> |
| Precision                                                                    | 0.6238 (±0.0132)        | 0.6068 (±0.0274)        | 0.6416 (±0.0144)        | 0.7717 (±0.0529)        | 0.7325 (±0.0239)        | 0.7595 (±0.0440)        | <b>0.8121 (±0.0550)</b> |
| Recall                                                                       | 0.7016 (±0.0465)        | <b>0.8835 (±0.0566)</b> | 0.7790 (±0.0379)        | 0.8051 (±0.0581)        | 0.7597 (±0.0105)        | 0.7920 (±0.0345)        | 0.8242 (±0.0586)        |
| F1 Score                                                                     | 0.6601 (±0.0276)        | 0.7194 (±0.0380)        | 0.7034 (±0.0224)        | 0.7873 (±0.0505)        | 0.7455 (±0.0082)        | 0.7748 (±0.0335)        | <b>0.8143 (±0.0102)</b> |