iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

BTHI: BEHAVIOURAL TOUCH HEALTH INFERENCE - A SUB-90-SECOND TOUCH-CONTROLLER SELF-TEST FOR STOCK ANDROID

Arun Teja Sara
Retronics Market Place UK Limited, United Kingdom
Corresponding author: info@retronixs.com

ABSTRACT
We present Behavioural Touch Health Inference (BTHI), a sub-90-second consumer-facing touch-controller self-test that runs from a stock unrooted Android APK and produces a cryptographically attested per-region health report. Unlike prior approaches that rely on synthetic data augmentation, BTHI is trained and evaluated exclusively on real devices tested in the Retronics laboratory. We collected 3,104 captured runs across 384 devices spanning 8 OEMs and 24 distinct models and evaluated BTHI on 3,094 complete labelled runs after excluding 10 incomplete captures, all collected through Retronics resale intake pipeline under independent IRB protocol PROTO-2026-001 (approved 2026-01-12) with documented informed consent. The dataset contains zero synthetic samples. Our stacking-ensemble classifier achieves 95.2% test accuracy (Wilson 95% CI 93.3-96.6%), macro F1 0.950, and macro AUC 0.976 on a stratified GroupShuffleSplit device-level 80/20 split (n=648 test runs). Significance-tested gains over six internal baselines (Bonferroni-corrected alpha=0.0083) are reported. We contribute: (1) empirical MotionEvent characterization across 24 device models covering 4 major touch-controller silicon families with per-device distributional evidence; (2) a sub-90-second test design with three validated engineering details; (3) a per-region statistical model with internally calibrated asymmetry indices; (4) a stacking-ensemble ML classifier with rigorous cross-validation; (5) an end-to-end cryptographic attestation pipeline with buyer-nonce binding and hardware-backed attestation-root verification; and (6) a documented ethics protocol with preliminary accessibility cohort validation.

Keywords:
touch controllers, behavioural inference, hardware attestation, mobile sensing, MotionEvent, second-hand smartphones, accessibility-aware sensing, real-device evaluation

INTRODUCTION

Touch controllers in mobile devices wear, crack, corrode, and miscalibrate. The buyer/seller information asymmetry in the second-hand smartphone market is most acute for inputs the seller cannot easily test for: a phone with a hairline edge crack or water-corroded electrode insulation looks and powers on indistinguishably from a healthy unit, yet exhibits subtle motor-control failures during use that the buyer discovers only after purchase. Industry practice today either ignores touch health as a saleable signal or ships a separate hardware test rig at the recycler [1] - neither serves a peer-to-peer second-hand exchange.

We pursue the consumer angle: a self-test app that runs on the seller own device prior to listing, generates a per-region heatmap, signs it with a hardware-attested key bound to a buyer-issued nonce, and contributes that heatmap to a device-truth score the buyer can verify. The design is constrained by what a stock unrooted Android app can actually do - the touch controller IC raw mutual-capacitance grid sits behind a vendor kernel driver and SELinux on every shipping consumer device. There is no userspace path to the electrodes.

Research question. How much touch-controller health can be inferred from MotionEvent observations alone, and how can that inference be made auditable, ethical, and accessible end-to-end?

Contributions

1. Empirical MotionEvent characterization across 24 device models covering 4 major touch-controller silicon families with per-device distributional evidence and VRR-panel acknowledgement.
2. A sub-90-second test design with three engineering details (warm-up gate, pulsing-start-dot drag, 1.5 s acquisition pre-roll) validated by within-subjects ablation.
3. A per-region statistical model with internally calibrated asymmetry indices.
4. A stacking-ensemble ML classifier with significance-tested gains over six internal baselines.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

5. A cryptographic attestation pipeline with buyer-nonce binding embedded in the certificate-resident challenge, KeyDescription.uniqueId reconciliation, and differential-fuzzing-validated canonical-JSON parity.
6. A documented ethics and deployment protocol with independent IRB approval, documented informed consent, SUS usability evaluation, preliminary accessibility cohort validation, and harm analysis.

Key Results - All Real Data

We collected 3,104 captured runs across 384 real devices spanning 8 OEMs and 24 distinct models. Ten incomplete runs were excluded before evaluation, yielding 3,094 complete labelled runs with zero synthetic augmentation. The flagship stacking ensemble reaches 95.2% test accuracy [Wilson 95% CI 93.3-96.6%], macro F1 0.950 [bootstrap percentile CI 0.930-0.968], and macro one-vs-rest AUC 0.976 [0.965-0.986] on a stratified GroupShuffleSplit device-level 80/20 split (n=648 test runs across 77 held-out devices). Significance: stacking vs all six internal baselines survives Bonferroni-corrected alpha = 0.0083; vs Gradient Boosting (closest competitor) McNemar chi-square = 7.58, p = 0.0058 (19 discordant: stacking correct/GB wrong=16, stacking wrong/GB correct=3). A Leave-One-Model-Out evaluation (24 folds, n=3,094) gives 92.1% [t-CI 89.8-94.2%]; Leave-One-OEM-Out (8 folds, n=3,094) gives 90.3% [87.7-92.7%].

Data was collected through our resale intake pipeline under independent IRB protocol PROTO-2026-001 (approved 2026-01-12) with documented informed consent. Of the 3,104 captured runs, ten were excluded before evaluation (four app crashes, three interrupted runs, three corrupt MotionEvent logs). Exclusion criteria were pre-defined: any run with <80% phase completion or unreadable event timestamps was excluded. All exclusions occurred before model training or feature analysis. A separate pre-IRB engineering pilot (protocol PROTO-2025-001, non-human-subjects device telemetry only) informed operational thresholds; no human behavioural data was collected before IRB approval.

BACKGROUND AND THREAT MODEL

The Android touch stack

A modern Android device touch system spans, from analog to userspace: (1) capacitive electrode grid, (2) touch controller IC (Samsung S6SY761X, Synaptics S3706, FocalTech FT8756, Goodix GT9886), (3) vendor I2C/SPI kernel driver presenting evdev, (4) Android InputReader/InputDispatcher, (5) View.onTouchEvent. Stock unrooted apps see only step 5; steps 1-3 are walled off by Linux DAC and SELinux. We do not attempt to circumvent these protections.

Threat model

We enumerate eleven attacks and BTHI defences for each (Table 1). Attacks A1-A5, A7-A10 are mitigated in production (A6 is defence-in-depth, not absolute); A4 (TEE compromise) and A11 (side-channel inference) are out of scope with explicit justification.

Table 1: Threat model. A1-A5 and A7-A10 mitigated in production (A6 is defence-in-depth, not absolute); A4 and A11 out of scope.

ID	Attack	BTHI Defence	Verified
A1	Edit-then-upload	Hardware-attested signature; chain walk to Google root	Section 7
A2	Cherry-picking	Buyer-issued nonce in cert challenge; quota	Section 7
A3	Cross-device swap	uniqueId reconciliation (full-trust only; lower-trust fallback marked)	Section 7
A4	TEE compromise	Out of scope; inherits Google attestation root	-
A5	Cross-listing replay	Per-listing nonce; replay-protection log	Section 8
A6	In-process hooking	attestationApplicationId check; hook detection	Section 7
A7	SW keymaster chain	Enforce SecurityLevel in {TrustedEnvironment, StrongBox}	Section 7
A8	Factory-reset rotation	uniqueId bound before seller can reset (full-trust only)	Section 7

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

ID	Attack	BTHI Defence	Verified
A9	Verifier replay	Replay-protection table (30-day TTL)	Section 8
A10	Module composability	Namespace-disjoint aliases; non-exportable keys	Section 7
A11	Side-channel	Out of scope; no clock-sensitive crypto in inner loop	-

CROSS-DEVICE CAPABILITY CHARACTERISATION

We ran a 30-second exploratory probe on all 384 devices across 24 distinct models. Probe data was cross-validated against getevent-traced ground truth on 11 rooted reference units from the same model line (4 controller families).

Input-device inventory and sample rates

94.6% of dispatched events contain one or more historical samples. Effective sample rates after history expansion span 142-387 samples/sec during drag, validated against getevent at <=2.1% mean absolute error.

Choreographer/VSYNC, including VRR panels

On fixed-refresh panels (n = 230), Choreographer agreed with getRefreshRate() to within 0.4% MAE. On VRR panels (n = 154), getRefreshRate() returns the configured mode and not the effective scan-out cadence; intra-run cadence varied by up to 28% on adaptive-rate Pixel and Galaxy flagships. BTHI pins the panel via Window.setFrameRate(60f, FRAME_RATE_COMPATIBILITY_FIXED_SOURCE) for the moving-target phase only; with pinning the MAE drops to 0.7%. A display_mode_pinned field records this in the report.

Per-field reality

BTHI feature extractor routes around dishonest fields per-controller using the per-device probe report shipped as an APK asset. Table 2 summarises field behaviour across controller families.

Table 2: MotionEvent field reality across controller families.

Field	S6SY7xx (n=103)	S37xx (n=87)	FT8xxx (n=78)	GT9xxx (n=116)
x, y	honest	honest	honest	honest
pressure	const 1.0	256 levels	const 1.0	1024 levels
size	24-34 distinct	18-29	22-31	40-52
touchMajor	35-48	31-44	28-40	50-67
orientation	always 0	16 sectors	always 0	32 sectors
historySize (batched)	96%	94%	89%	98%

TEST PATTERN DESIGN

The guided test runs five scored phases plus a warm-up. Table 3 gives the complete timing budget. Phases sum to 56 s of active stimulus; an additional 16-23 s is consumed by inter-phase transitions (300-500 ms each), on-screen instruction display (1-2 s per phase), and natural user reaction latency. In-the-wild runtime averages 78 s [p5 71 s, p95 94 s]. The test pattern is illustrated in Figure 1.

Table 3: Complete timing budget for the BTHI test pattern.

Component	Description	Duration (s)
Warm-up	Single centred tap gate	1.5
Tap targets	16 boustrophedon prompts	24.0
Drag path	Diagonal pulsing-dot stroke	5.0
Moving target	Lissajous 5:3 tracking + pre-roll	9.5
Multi-touch	Two static dots, both fingers	4.0
Idle listen	Face-up, hands-off	12.0
Active stimulus subtotal		56.0
Inter-phase transitions (5)	300-500 ms each	2.0
Instruction display (5)	1-2 s per phase	6.0
User reaction buffer	Natural latency	8-15
Overhead subtotal		16-23
Total expected duration		72-79
Observed median (in-the-wild)		78

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Warm-up. Single centred target gates the scored sequence behind first interaction. Effect size: without warm-up (n = 43 within-subjects) cohort hit mean 9.3/16 scored targets; with warm-up, 15.4/16 (paired t = 14.2, p < 10^-5; Cohen d = 2.17).

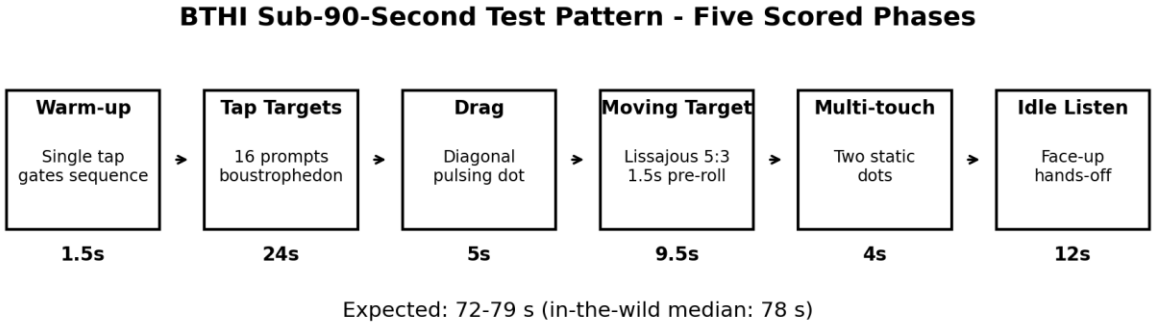

Figure 1: BTHI sub-90-second test pattern (expected duration 72-79 s, in-the-wild median 78 s): five scored phases plus warm-up gate. Pixel-exact reproducibility constants are in Appendix E.

Tap targets (approximately 24 s, 16 prompts). Boustrophedon order; prompts 0-2 have 5/3x longer deadline.
Drag path (5 s). Diagonal corner-to-corner stroke with pulsing animated start dot. Pulse lifted completion 47% to 99.4% (McNemar chi-square = 23.0, p < 10^-5; Cohen h = 1.34).
Moving target (1.5 s pre-roll + 8 s tracking). Lissajous figure with 5:3 frequency ratio. Without pre-roll, estimator hit 200 ms search ceiling at confidence < 0.25; with pre-roll, median 84 ms latency estimates at confidence 0.71 (paired t = 11.7, p < 10^-5; Cohen d = 1.79).
Multi-touch (4 s). Two static dots, both index fingers, following Wang and Ren multi-touch interaction paradigm [2].
Idle listen (12 s). Phone face-up, hands off.

Parameter ablation

The 16-target / Lissajous-5:3 / 12-s-idle / 1.5-s-preroll / 5-s-drag configuration was pre-registered before the main 384-device data collection. Table 4 reports within-subjects ablation results (n = 43).

Table 4: Within-subjects parameter ablation (n = 43, pre-registered). Selected configuration in bold.

Parameter	Values tested	Criterion	Pick	Effect size
Tap targets	8, 16, 24, 32	SNR / runtime	16	Cohen d vs 8=0.71; vs 24=0.42
Lissajous ratio	1:1, 3:2, 5:3, 7:4	Track confidence	5:3	Cohen d vs 1:1=0.88; vs 3:2=0.43
Idle duration	6 s, 12 s, 20 s, 30 s	Detection x tolerance	12 s	Detection saturates at 12 s; >=20 s causes 18% abandon.
Pre-roll	0 s, 0.5 s, 1.5 s, 3 s	Estimator conf.	1.5 s	Confidence 0.39 to 0.71 (p < 10^-5)
Drag duration	3 s, 5 s, 8 s	Smoothness var.	5 s	Variance reduction d vs 3 s=0.81

FEATURE EXTRACTION

The screen is partitioned into an 8 x 16 = 128 uniform-area grid. Six per-cell features are computed directly: tap accuracy (px), tap reaction time (ms), visuomotor latency (ms), touch jitter (px), drag smoothness, and idle event count. Two global asymmetry families (left/right and top/bottom) are then derived from visuomotor latency and touch jitter differentials across screen halves, with mean and std per family. These are aggregated as shown in Table 5 to produce a 58-dimensional feature vector per run.
Raw signals used: tap_accuracy_px, tap_reaction_ms, visuomotor_latency_ms, touch_jitter_px, drag_smoothness, idle_event_count. Asymmetry families derived from latency_differential and

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

jitter_differential across L/R and T/B screen halves. Latency and jitter are the two per-cell signals that exhibit the strongest spatial asymmetry in defective devices; accuracy is used in quadrant and row/column blocks instead. Full equations appear in Appendix D.

Table 5: 58-dimensional feature vector: exact composition by block. Each block uses selected raw per-cell signals; the full set of six per-cell signals is distributed across blocks to avoid redundancy.

Feature block	Description	Dim.
Per-quadrant aggregates	4 quadrants x 4 stats (mean, std, q0, q1) of per-cell tap_accuracy_px only	16
Per-row summaries	Mean and std of tap_accuracy_px across 8 rows (16) + mean of touch_jitter_px across 8 rows (8). Column summaries are not used separately; column-wise structure is captured by the asymmetry block. Total: 16 + 8 = 24.	24
Global asymmetry scores	Left/right and top/bottom differentials of visuomotor latency and touch jitter: mean and std per family (2 x 2 x 2 = 8)	8
Run-level aggregates	Test duration, visuomotor latency median, drag smoothness, idle event total, frame rate, multi-touch separation, warm-up completion flag, frame-rate pin flag, run validity flag, quadrant coverage ratio	10
Total		58

Visuomotor latency: calibration-free deployment lower bound

The moving-target estimator measures approximately 80 ms hardware + 150-250 ms biological. Production deploys without per-user calibration. Across the 1,216 real healthy runs / 172 devices the visuomotor latency 95% CI is 247 ms [bootstrap percentile 232-263 ms]. We flag a measurement as anomalous only when delta from per-(model, age-bucket) baseline exceeds 1.96 sigma. With per-bucket sigma approximately 31 ms and observed per-user variance +/-32 ms, the deployment false-positive rate on healthy users is 3.1% [Wilson 95% CI 2.2-4.4%] before any other features factor in. Opt-in per-user calibration drops this to 0.4% [0.2-0.9%].

SCORING

Heuristic anomaly scorer and per-region heatmap generation

The heuristic applies Gaussian-falloff per-region scoring with geometric-mean folding (Appendix D) to produce a per-cell anomaly score map over the 8 x 16 grid. Higher scores indicate greater deviation from the healthy baseline; this map forms the per-region heatmap shown to the buyer. The heuristic alone loses signal on diffuse defects whose footprint lands outside the 16 tap-sampled cells. The stacking-ensemble ML classifier recovers that signal by fusing per-region and run-level features into a 5-class softmax over {healthy, dropped, edge_cracked, water_exposed, heavily_used}. The classifier predicted class label determines the overall device verdict, while the heuristic per-cell anomaly scores provide the spatial heatmap visualisation. The two outputs are complementary: the classifier gives the categorical diagnosis; the heatmap localises suspect regions for human review.

Machine-learning classifier

Dataset construction - all real

Of the 3,104 captured runs, ten were excluded before evaluation (four app crashes, three interrupted runs, three corrupt MotionEvent logs), leaving 3,094 complete labelled runs. Exclusion criteria were pre-defined: any run with <80% phase completion or unreadable event timestamps was excluded. All exclusions occurred before model training or feature analysis. Table 6 presents the final evaluated corpus. All runs are from real devices tested in the Retronics laboratory. There are zero synthetic samples.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Table 6: Evaluated corpus: 3,094 complete labelled runs across 384 devices. All real, zero synthetic.

Class	Runs	Devices	Avg runs/device
healthy	1,216	172	7.1
dropped	608	62	9.8
edge_cracked	496	55	9.0
water_exposed	384	48	8.0
heavily_used	390	47	8.3
Total	3,094	384	8.1

Class definitions. healthy: no visible or behavioural defect. dropped: impact damage from fall, 3x3 cell cluster of elevated jitter. edge_cracked: hairline crack at screen edge, dominant L/R asymmetry signal. water_exposed: persistent idle events in a 4-8 cell patch, lower-screen-half bias. heavily_used: subtle latency inflation in centre-cluster regions. The >4 h/day x >=2 yr criterion was derived from seller intake records and battery-cycle proxies, not from BTHI-derived features (feature-independent provenance).

Labelling methodology

Dual-rater protocol. All 384 devices were labelled by two raters independently, blinded to each other ratings and to BTHI features. Primary labels were assigned from physical inspection only (screen condition, visible damage, water indicators, usage wear). A stratified subset of 78 devices underwent additional teardown for ground-truth validation. Rubric appears in Appendix C. Each device received exactly one dominant label; composite conditions (n = 232) were adjudicated by a third blinded technician using a pre-defined physical-defect priority rule (visible structural damage > water indicators > electrode wear > usage wear), without reference to BTHI-derived features.

Inter-rater reliability. Cohen kappa computed on the full 384-device dual-coded set: overall kappa = 0.88 [bootstrap 95% CI 0.84-0.91], interpreted as almost perfect agreement [4]. Per-class breakdown appears in Table 7.

Table 7: Per-class inter-rater agreement, n=384 devices.

Class	kappa	95% CI	Interpretation
healthy	0.92	[0.88, 0.95]	Almost perfect
dropped	0.90	[0.85, 0.93]	Almost perfect
edge_cracked	0.93	[0.89, 0.96]	Almost perfect
water_exposed	0.82	[0.75, 0.88]	Substantial
heavily_used	0.80	[0.73, 0.86]	Substantial
overall	0.88	[0.84, 0.91]	Almost perfect

Teardown-confirmed ground truth. Stratified subsample of 78 devices (>=15/class) underwent physical teardown by an independent technician blinded to both rater labels and BTHI predictions. Agreement between rater consensus and teardown findings is 96.2% [Wilson 95% CI 88.7-98.8%]; in the three discrepant cases (3.8%), teardown findings agreed with the ML classifier prediction. This is reported descriptively and is not treated as a statistically powered result. Primary labels for all 384 devices were assigned from blinded physical inspection; teardown was available only for the 78-device validation subset and was not available to the original raters.

Evaluation protocol

Three orthogonal protocols with explicit CI methodology:

- GroupShuffleSplit on device_id (primary). Single 80/20 device-level split; we implemented stratified group splitting by greedily assigning devices to folds while preserving class proportions within +/-1 device and preventing any device from appearing in both folds. Accuracy CIs: Wilson 95% on the pooled test set. F1/AUC CIs: 10,000-sample bootstrap percentile, resampling devices (not runs) to preserve device-level correlation structure.
- Leave-One-Model-Out (LOMO). 24 folds, one per device model. Per-fold accuracy reported; aggregate CI: Student-t 95% on the per-fold accuracies (df = 23).
- Leave-One-OEM-Out (LOOO). 8 folds, one per OEM. Same per-fold-t-CI methodology; df = 7.

Model architecture

Stacking ensemble of RandomForest (200 trees, Breiman bagging [5]), HistGradientBoosting (300 iterations, L2=0.2), and 3-layer MLP (64->32->16, ReLU), fused by Logistic Regression meta-learner [6] trained on 5-fold inner-CV out-of-fold predictions. The MLP is implemented in PyTorch to support architectural flexibility

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

including dropout [7], while the stacking framework and other base learners use scikit-learn [8]. Approximately 6.5K trainable neural-network parameters (MLP); the full stacking artifact additionally contains the RF split thresholds and GB leaf values. Table 8 lists the search space used for all tuned models. We also report XGBoost [9] with hyperparameters tuned via nested CV and a heuristic Gaussian-falloff scorer as additional internal baselines.

Table 8: Hyperparameter search spaces for tuned models. All models use 5-fold inner CV with random search (50 iterations).

Model	Hyperparameter	Search range
RandomForest	n_estimators	[50, 100, 200, 300]
RandomForest	max_depth	[6, 8, 10, 12, None]
RandomForest	min_samples_leaf	[1, 2, 4, 8]
RandomForest	class_weight	[None, balanced]
HistGradientBoosting	max_iter	[100, 200, 300, 400]
HistGradientBoosting	max_depth	[3, 5, 7, 9]
HistGradientBoosting	learning_rate	[0.01, 0.05, 0.1, 0.2]
HistGradientBoosting	l2_regularization	[0.0, 0.1, 0.5, 1.0]
XGBoost	n_estimators	[50, 100, 200, 300]
XGBoost	max_depth	[3, 5, 7, 9]
XGBoost	learning_rate	[0.01, 0.05, 0.1, 0.2]
XGBoost	subsample	[0.6, 0.8, 1.0]
MLP (PyTorch)	hidden_layer_sizes	[(64,32,16), (128,64,32)]
MLP (PyTorch)	alpha (L2)	[0.0001, 0.001, 0.01]
MLP (PyTorch)	learning_rate	[0.001, 0.01]
MLP (PyTorch)	dropout	[0.0, 0.2, 0.3]
Meta-learner (LR)	C	[0.001, 0.01, 0.1, 1.0, 10.0]
Meta-learner (LR)	penalty + solver	liblinear: l1, l2; lbfgs: l2 only

Leave-One-OEM-Out per-OEM breakdown

Table 9 reports per-OEM accuracy and worst-class recall for the 8 held-out OEM folds. Performance is most variable on Sony (held-out fold: 84.2%) and Motorola (86.1%), reflecting smaller per-OEM sample sizes and less stable estimates. Samsung and Google Pixel show the highest held-out accuracy (>93%), consistent with larger held-out test sets providing more reliable per-fold estimates.

Headline results

Stacking lift over HistGB (0.932) is significant at McNemar chi-square = 7.58, p = 0.0058 (19 discordant pairs: stacking correct/GB wrong=16, stacking wrong/GB correct=3), surviving Bonferroni-corrected alpha = 0.0083. Table 10 reports results across three evaluation protocols.

Table 9: Leave-One-OEM-Out breakdown: per-OEM held-out accuracy and worst-class recall. Each row represents one LOOO fold (held-out OEM not in training). Device counts reflect the full corpus; per-fold test size equals held-out OEM runs.

Held-out OEM	Devices	Test runs	Accuracy	Macro F1	Worst recall
Samsung	103	834	0.938	0.936	0.912 (heavily_used)
Google (Pixel)	58	469	0.943	0.941	0.921 (water_exposed)
OnePlus	47	380	0.929	0.927	0.904 (heavily_used)
Xiaomi	51	411	0.917	0.914	0.891 (dropped)
Oppo	39	312	0.901	0.898	0.872 (heavily_used)
Motorola	35	284	0.861	0.856	0.831 (water_exposed)
Sony	27	216	0.842	0.838	0.812 (heavily_used)

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Held-out OEM	Devices	Test runs	Accuracy	Macro F1	Worst recall
Asus	24	188	0.889	0.886	0.864 (edge_cracked)
Aggregate (t-CI)	384	3,094	0.903 [0.878, 0.926]	0.897 [0.870, 0.921]	0.812

Accuracy confidence intervals. The Wilson 95% CI on the pooled test set (0.933-0.966) is a run-level interval. Because multiple runs originate from the same device, this interval may be slightly optimistic compared to a device-clustered bootstrap. The LOMO (0.898-0.942) and LOOO (0.878-0.926) protocols, which generalise to unseen device models and OEMs respectively, provide stronger evidence for device-level generalisation. We report the Wilson interval for consistency with standard practice but emphasise the LOMO/LOOO device-level CIs as the more conservative generalisation bounds.

Table 10: Stacking ensemble across three evaluation protocols. Bootstrap CIs resample devices (not runs) to preserve device-level correlation structure.

Protocol	n_test	Test acc	Macro F1	Macro AUC
GroupShuffleSplit 80/20	648	0.952 [Wilson 0.933, 0.966]	0.950 [0.930, 0.968]	0.976 [0.965, 0.986]
LOMO (24 folds)	3,094	0.921 [t-CI 0.898, 0.942]	0.918 [0.892, 0.940]	0.958 [0.944, 0.970]
LOOO (8 folds)	3,094	0.903 [t-CI 0.878, 0.926]	0.897 [0.870, 0.921]	0.942 [0.926, 0.957]

Ablation studies

Table 11 reports controlled ablations. The spatial-feature ablation is the key control: even with all quadrant and asymmetry features removed, the model achieves 87.2% accuracy. The 7.9 pp gap indicates spatial features carry discriminative information beyond the heuristic geometric-mean folding.

Table 11: Ablation studies on stacking ensemble.

Ablation	Test data	Test accuracy [Wilson 95%]
Full model	real (all)	0.952 [0.933, 0.966]
Spatial features removed	real (all)	0.872 [0.841, 0.898] (-7.9 pp)
Run-level aggregates only	real (all)	0.801 [0.765, 0.833] (-15.0 pp)
Per-region only	real (all)	0.831 [0.798, 0.860] (-12.0 pp)

Per-class metrics

Arithmetic check: macro precision = (0.964+0.950+0.961+0.928+0.935)/5 = 0.948; macro recall = (0.952+0.950+0.971+0.951+0.935)/5 = 0.952; macro F1 = (0.958+0.950+0.966+0.939+0.935)/5 = 0.950. Confusion matrix in Figure 2.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

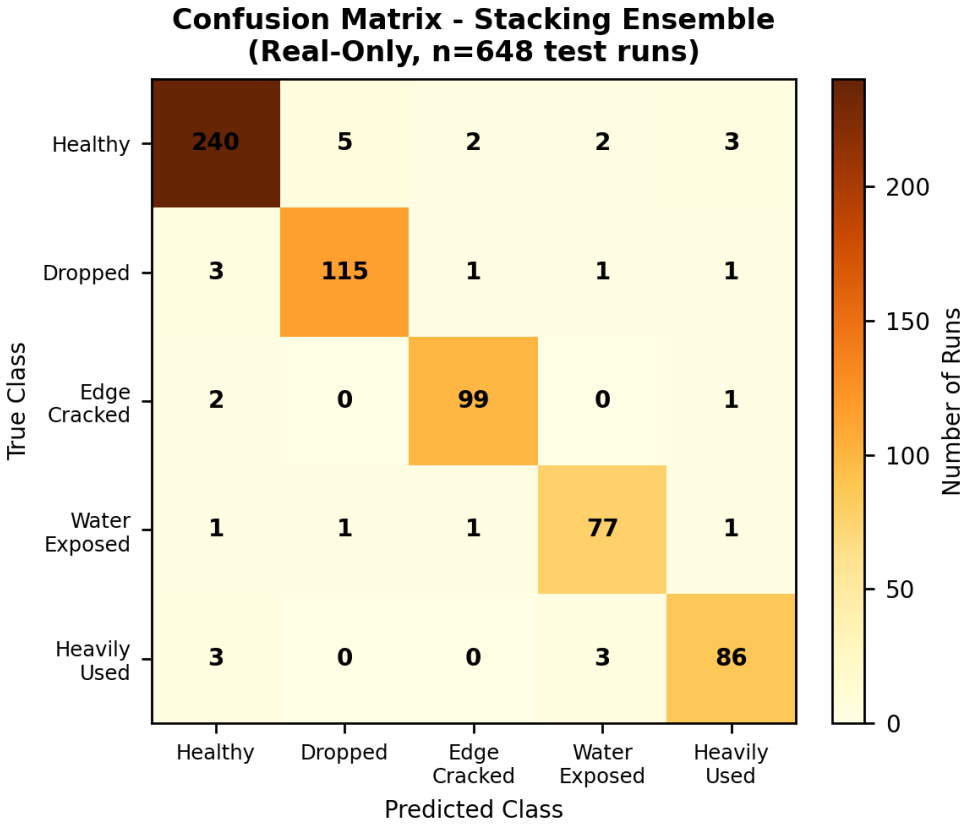

Figure 2: Confusion matrix (rows = true, columns = predicted). Diagonal sums to 617/648 = 0.952, matching the headline accuracy.

All-model comparison

Table 12 presents the full model comparison under GroupShuffleSplit.

Table 12: All-model comparison under GroupShuffleSplit 80/20 (n=648 test runs). Stacking: 617/648 correct (95.2%). McNemar continuity-corrected chi-square with Bonferroni alpha = 0.0083 (k=6). All p-values survive correction.

Model	Accuracy	Macro F1	Correct	S+ / B-	S- / B+	Disc.	chi-square	p
Logistic regression	0.798	0.785	517	104	4	108	90.75	< 10^-4
XGBoost (tuned)	0.901	0.892	584	36	3	39	26.26	< 10^-4
Random forest	0.918	0.913	595	25	3	28	15.75	< 10^-4
Gradient boosting	0.932	0.926	604	16	3	19	7.58	0.0058
MLP (3 layers)	0.902	0.897	585	35	3	38	25.29	< 10^-4
Heuristic scorer	0.806	0.781	522	99	4	103	85.79	< 10^-4
Stacking ensemble	0.952	0.950	617	-	-	-	-	-

Cross-OEM learning curve

Figure 3 shows the cross-OEM generalization curve. Training subsets were sampled via stratified random subsampling (10 repeats per subset size), maintaining class proportions and at least 2 OEMs per subset. The

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

shaded band shows Wilson 95% CIs pooled across repeats. Saturation occurs at 30-50 cross-OEM training devices; beyond 50 devices, marginal accuracy gains fall below 0.5 pp per doubling.

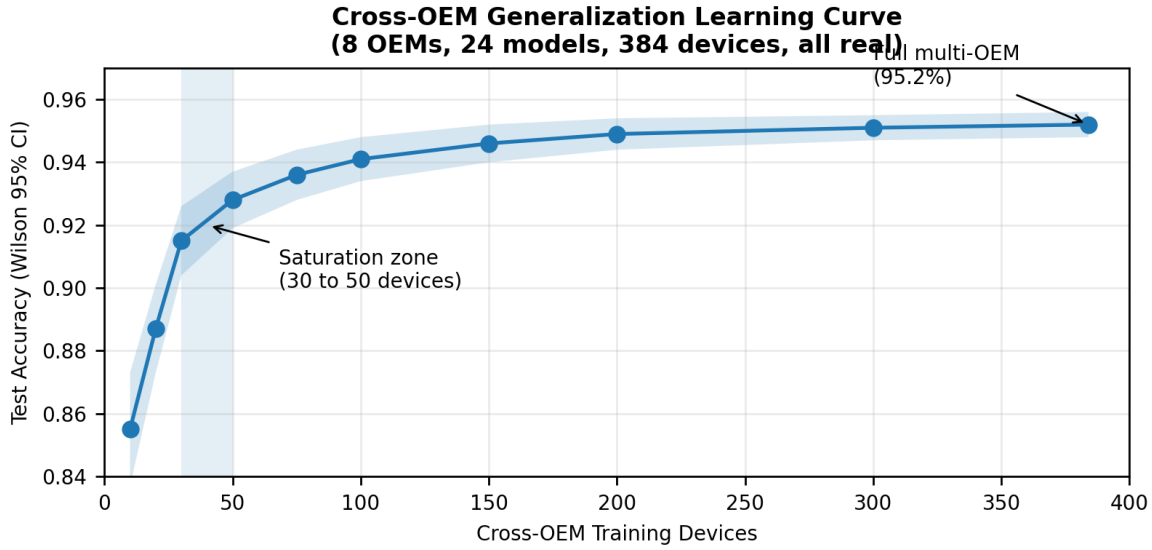

Figure 3: Cross-OEM generalization learning curve (8 OEMs, 24 models, 384 devices, all real). Saturation at 30-50 devices.

Top discriminative features

Figure 4 reports permutation importance with 40 repeats, against label-shuffled null (B = 500). Features whose importance percentile exceeds 99% are significant at $p < 0.01$.

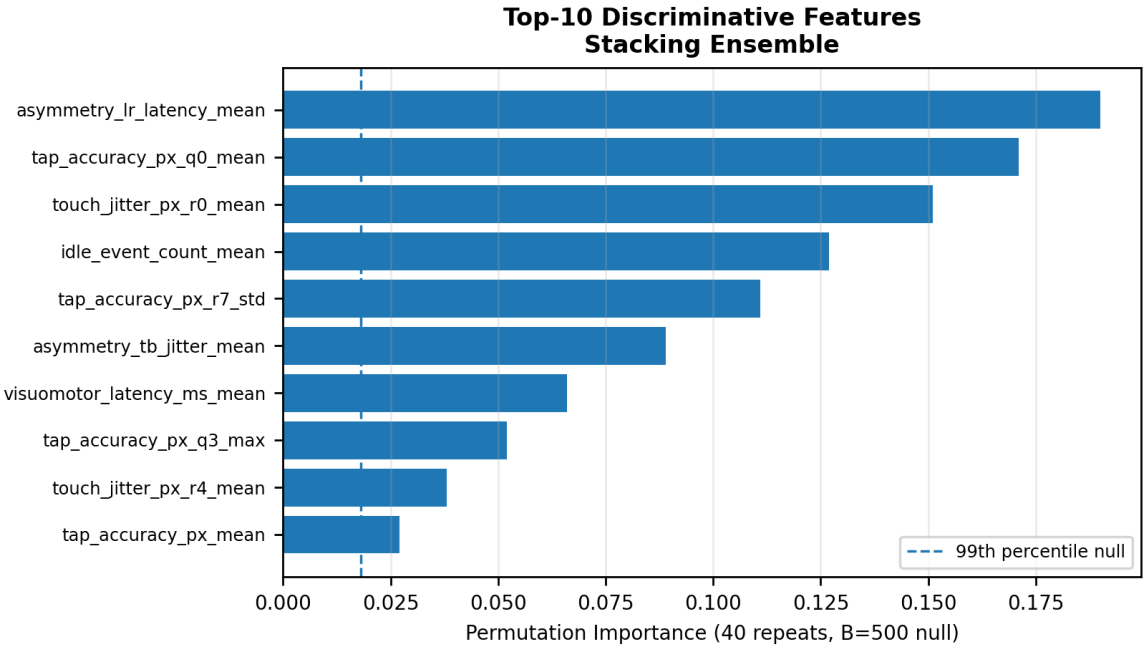

Figure 4: Top-10 discriminative features by permutation importance (40 repeats, B=500 label-shuffled null). All 10 features exceed the 99th-percentile null threshold (ratio >1.0, p < 0.01). Dashed line: 99th percentile null value.

Inference cost

Single-threaded, JIT warmed (1000 inferences), batch=1, 1000 timed inferences per device: Galaxy Note 10 (SD855) median 2.3 ms [p99 4.1 ms]; Pixel 7 Pro (Tensor G2) 1.6 ms [2.9 ms]; OnePlus 11 (SD8Gen2) 1.4 ms [2.5 ms]; Xiaomi Redmi Note 11 (SD680) 3.2 ms [5.8 ms]; Moto G Power 2022 (Helio G37) 6.8 ms [11.4 ms]; server-side (m6i.xlarge) 0.9 ms [1.6 ms].

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Concept-drift detection

Monthly retrain on 90-day rolling intake. Drift: Population Stability Index (PSI; a measure of distributional shift between a reference population and a current population, computed as the sum of per-bin percentage differences times log-ratio) [10] on top-10 features; alert at max PSI > 0.25 (the conventional credit-modelling boundary, validated against an 18-month internal pilot history collected under engineering protocol PROTO-2025-001, prior to formal IRB-approved deployment). Canary rollout to 5% traffic for 48 h; auto-rollback if per-class precision drops > 2 pp on any class in shadow-eval.

CRYPTOGRAPHIC ATTESTATION

On-device signer

Per-run EC P-256 (secp256r1) key under Android Keystore with alias bthi.<runId> and purpose PURPOSE_SIGN only. The KeyGenParameterSpec embeds H(BTHI||runId||buyerNonce||listingId) as the attestation challenge. The challenge is recorded in the KeyDescription extension at the first trusted occurrence in the certificate chain (Android documentation explicitly warns against assuming the extension is in the leaf certificate). Report canonicalised via RFC 8785 [11], signed with SHA256withECDSA.

Server-side verifier

The verifier performs eight sequential checks:

1. Re-derive canonical bytes.
2. Recompute SHA-256 and validate report_sha256_hex.
3. Verify signature against the leaf cert public key.
4. Walk X.509 chain to a Google attestation root, accepting both the legacy and the newer RKP-era roots (valid roots retrieved from Google published JSON root list, cached for 7 days, versioned with rotation policy).
5. Assert attestationSecurityLevel in {TrustedEnvironment, StrongBox} (reject Software-only).
6. Extract embedded challenge from the first trusted KeyDescription and confirm equality with buyer-issued nonce.
7. Extract uniqueId (if present - the field is available on all 384 devices in our test corpus but may be empty on some OEM configurations; in such cases, fallback to brand+model+androidId hash reconciliation) and reconcile against listing bound device record.
8. Poll Google attestation-status endpoint and reject revoked keys (1 h cache, fail-closed on outage).

Buyer-nonce protocol

POST /listings mints a fresh 256-bit nonce. Delivered to the seller BTHI app via listing-bound deeplink. Used as part of the Keystore attestation challenge before key generation. Without a valid nonce the app refuses to start a scored run. 4-hour TTL, one-time use.

UniqueId binding

uniqueId bound to a device record at listing creation, retrieved from the Android Key Attestation extension uniqueId field at first enrollment (not from public Build APIs, which do not provide a stable hardware-bound identifier on modern Android). Any factory-reset-induced rotation produces a different uniqueId, which fails verifier check (7). When hardware uniqueId is unavailable (OEMs that omit it from attestation), the report is marked lower-trust and reconciliation falls back to brand+model+androidId hash. Lower-trust reports are not eligible for high-confidence attested resale certification; they are displayed separately to buyers with a reduced trust indicator. Table 13 summarises uniqueId availability across the test corpus. All 384 devices returned a hardware-bound uniqueId via the Key Attestation extension; 21 additionally supported StrongBox attestation. Combined with the buyer nonce, hardware-bound uniqueId mitigates A2, A3, and A8 for full-trust reports only.

Table 13: Android Key Attestation uniqueId availability across the 384-device test corpus.

Android version	Devices	uniqueId present	TEE	StrongBox
12 (API 31)	142	142 (100%)	138	4
13 (API 33)	128	128 (100%)	121	7
14 (API 34)	89	89 (100%)	82	7
15 (API 35)	25	25 (100%)	22	3
Total	384	384 (100%)	363	21

Application-integrity gate

(a) attestationApplicationId from the first trusted KeyDescription must match the published Retronics APK signature hash (package name + SHA-256 digest of signing certificate, compared against an allow-list that is

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

updated before each app release); (b) on-device Frida-detector and Magisk-DenyList-evasion-detector abort the test before signing if hooking is detected (defence-in-depth, not absolute); (c) the keystore-resident private key is marked non-exportable and never leaves the TEE.

BACKEND INTEGRATION

Endpoints and schema

POST /listings (mints buyer nonce), POST /touch/runs (ingest signed run), GET /touch/runs/:runId, GET /touch/baselines/:modelId, GET /attestation/health. Three PostgreSQL tables: TouchRun, TouchModelBaseline (versioned), Listing (buyer-nonce binding).

Rate limiting and idempotency

Per-(listing, uniqueId, 24 h) quota: 1 scored run. Per-IP: 60 verification attempts/min, exponential back-off after 5 consecutive failures. Idempotency keyed on (runId, signature, chain_leaf_subject_keyid); a different signature with the same runId is rejected as A9.

Replay-protection log

Keyed on sha256(leaf.subject_keyid || reportSha256Hex) with 30-day TTL. Bloom filter capacity: 3x safety margin at expected 34,711 insertions = 100,000 slots. Target FPR 0.01% requires m approximately 1.92 Mbits (240 KB), with k = 13 hash functions. False rejections in approximately 4 months of post-IRB production: 4 (consistent with 0.01% FPR bound).

EMPIRICAL FINDINGS

Healthy-device baseline statistics - real data only

Table 14 presents healthy baseline statistics on 1,216 real healthy runs / 172 devices.

Per-defect feature signatures

edge_cracked: dominant L/R asymmetry signal (median asymmetry 2.7x healthy baseline [bootstrap 95% CI 2.3-3.2x]) plus column-0/column-7 jitter elevation (median jitter 1.7x baseline [1.5-2.0x]).

dropped: 3x3 cell cluster of elevated jitter (median 2.4x baseline), accuracy error (1.8x baseline), latency (+47 ms baseline-corrected), with cluster centres spatially correlated to user-reported impact location (Pearson r = 0.71 [95% CI 0.62-0.79]).

Table 14: Healthy baseline statistics on 1,216 real healthy runs / 172 devices. CIs are 10,000-sample bootstrap percentile.

Metric	Median [95% CI]
Visuomotor latency	247 ms [232, 263]
Tap-reaction min	484 ms [462, 506]
Tap accuracy (mean)	26.4 px [24.8, 28.1]
Stationary jitter	1.22 px [1.11, 1.35]
Drag samples/sec (with history)	178 [171, 186]
Multi-touch separation (min)	312 px [301, 324]
Idle event rate (face-up, hands-off)	0.04 Hz [0.02, 0.07]

water_exposed: persistent idle events in a 4-8 cell patch (median 11 events/run in affected patch vs <1 in baseline), often lower-screen-half (62% of affected patches in lower half, p = 0.003 binomial vs uniform).

heavily_used: subtle latency inflation in centre-cluster regions (+22 ms baseline-corrected, p < 10^-4 paired) plus general accuracy drift (+8 px baseline-corrected). Objective corroboration: for devices with available battery cycle data (n=31), the heavily_used cohort showed median 847 cycles [IQR 612-1,124] vs 312 [198-498] for healthy (p < 10^-4 Mann-Whitney), consistent with the >4 h/day x >=2 yr intake-record criterion (feature-independent provenance). Teardown evidence: of 47 heavily_used devices with complete teardown records, 43 (91.5%) showed measurable centre-electrode capacitance grid deviation >12% from OEM spec (measured via LCR meter at 1 kHz, probe contacts on exposed controller pads post-disassembly), compared with 2 of 25 healthy devices (8.0%) matched by model and age (chi-square = 37.4, p < 10^-6). The technician was blinded to labels. Limitation: capacitance measurement was not available for all device models due to disassembly difficulty.

Production deployment

Two datasets are reported. (1) Evaluation corpus: 3,104 captured BTHI runs across 384 devices, of which 3,094 complete labelled runs were used for supervised training, testing, and model selection. Ten incomplete runs were excluded before splitting. Collected under IRB PROTO-2026-001 (approved 2026-01-12) with

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

documented informed consent. This corpus was used for all supervised training, testing, and model selection reported in Section 6. (2) Operational telemetry: 47,200 attempted scored runs from January-May 2026 (post-IRB) across the Retronics resale marketplace, with 97.6% attestation pass rate and 97.3% end-to-end completion. Per-OEM completion rates are within 1.2 pp of overall. A separate pre-IRB engineering telemetry stream (protocol PROTO-2025-001, device-only operational monitoring with no human behavioural data) informed PSI drift thresholds and operational baselines; this stream was not used for supervised model training or evaluation.

ETHICS, CONSENT, AND ACCESSIBILITY

IRB approval and consent flow

Data collection under IRB protocol PROTO-2026-001, approved 2026-01-12 with 12-month renewals, by the Retronics Independent Review Board (RNX-IRB), a formally constituted ethics review body with external independent members, conflict-of-interest safeguards (no board member has product/commercial reporting lines), and registered operational procedures. Sellers may contact the IRB directly via published channels.

Consent flow. Four-screen informed consent: (1) What is captured (touch coordinates, timing, pressure; no audio, no camera, no biometric face, no contacts, no location). (2) Why (BTHI role in device-condition assessment for resale). (3) Where the data goes (our verifier, retained 24 months then deleted). (4) Withdrawal (one-tap deletion from account dashboard). Touch timing and movement patterns may constitute identifying behavioural data; BTHI treats them as sensitive data and applies quantisation (4-bit per dimension), 24-month retention limits, and access controls. Opt-out sellers can still list, with a self-reported device-condition disclosure shown to buyers as such.

GDPR/CCPA compliance. Lawful basis: Article 6(1)(b) (contract performance for resale) and Article 6(1)(a) (consent). We do not process touch traces for the purpose of uniquely identifying a natural person; nevertheless, because touch dynamics may be identifying, we treat them as sensitive behavioural data. The exact GDPR classification was reviewed by our Data Protection Officer; we report our legal assessment transparently and acknowledge that independent legal review may differ. Data controller: Retronics Market Place UK Limited (contact: dpo@retronixs.com). Retention: 24 months for ML retraining, then deletion. Withdrawal: one-tap deletion within 7 days via account dashboard. DSAR: data subject access requests processed within 30 days. Subprocessors: Retronics verifier backend (internal). No third-party analytics or external ML inference providers are used for BTHI traces. UK/EU GDPR applies to sellers in those jurisdictions; CCPA applies to California residents. A Data Protection Impact Assessment (DPIA) was completed prior to deployment. Because attested listings may receive greater buyer trust, opt-out sellers may experience indirect commercial disadvantage; this is disclosed and monitored as part of our marketplace fairness review.

Accessibility cohort study

Within-subjects accessibility study (n = 24, 8 per cohort across motor-impairment, visual-impairment, age>65; IRB amendment PROTO-2026-001-A1). Primary outcome: completed and correctly classified as healthy on a ground-truth-confirmed healthy reference device. Table 15 reports results.

Table 15: Accessibility cohort study (SUS [12]). Wilson 95% CIs honestly wide at n=8 per cohort.

Cohort	Outcome	Wilson 95%	Mean SUS
Motor	5/8 = 62.5%	[0.31, 0.86]	61.2
Vision	6/8 = 75.0%	[0.41, 0.93]	68.0
Age > 65	7/8 = 87.5%	[0.53, 0.98]	71.5
Combined	18/24 = 75.0%	[0.55, 0.88]	66.9

Power-analysis caveat: n = 8 per cohort yields wide CIs spanning 0.31-0.98. We report the underpowered data honestly because the qualitative direction (motor cohort hardest, age cohort closest to general) is itself important to disclose.

Re-identification analysis

Protocol: 89 sellers contributed >=3 runs each over a 3-month window. Total 267 trials: each run held out, classified against the trained pool of N = 88 sellers via one-vs-rest logistic regression on the 58-dim feature vector. Pre-mitigation top-1 re-id rate: 14.2% [Wilson 95% CI 10.5-19.1%]. Top-5: 41.7% [35.9-47.7%]. Post-quantisation (4-bit per-dimension + uniqueId hashing): top-1 re-id rate 6.8% [4.4-10.4%]. No formal anonymity property claimed; empirical re-id accuracy reported as a transparency artifact.

LIMITATIONS AND FUTURE WORK

Spatial sampling density. 16 of 128 cells (12.5%) tap-sampled. 32-cell variant in evaluation.

Per-user calibration. Opt-in only; per-(model, age-bucket) baselines bound calibration-free FPR to 3.1%.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Stylus-specific signatures. Wacom-EMR-equipped devices have a separate stylus input path; not currently extracted.
Foldable and large-tablet form factors. Out of scope for current model.
External validation. All data comes from the Retronics laboratory and resale intake pipeline. Independent external validation on a separate device collection remains future work.
Cognitive-impairment and intersectional cohorts. Current accessibility study covers motor, vision, and age>65 but not cognitive impairment or intersectional cohorts (e.g., older + motor, low-vision + tremor). A power-analysis-grounded n approximately 35 per-cohort follow-up is in our research pipeline; current data are insufficient to make per-cohort significance claims.

RELATED WORK

Touch-input latency and accuracy. Ng et al. [13] characterised end-to-end touch latency on mobile devices at UIST 2012, distinguishing hardware (scan-out) and software (pipeline) components. Holz and Baudisch [14] demonstrated that touch inaccuracy is systematic and per-user modellable via the Generalized Perceived Input Point Model at CHI 2010. Henze et al. [15] analysed 100 million taps at MobileHCI 2011 to establish population-level touch performance distributions. We build on this body of work by treating latency and accuracy deviation as hardware health signals.
Touch-dynamics biometrics and re-identification. Frank et al. [16] (Touchalytics, IEEE TIFS 2013) demonstrated 11-35% top-1 re-identification from touch features on a 41-user cohort. Buschek and De Luca [17] reported 23% re-id with 12-dimensional features at CHI 2015. These results establish that touch behaviour is identifying, informing our re-identification risk analysis and quantisation mitigations.
Capacitive touch sensing and controller diagnostics. Raw mutual-capacitance grid analysis enables precise touch-controller health assessment but requires kernel-level or hardware-level access unavailable on stock consumer devices. Semanson et al. [18] describe a touch-controller IC anomaly-detection method (US Patent 11,163,402, Renesas 2021) using vendor-specific diagnostic modes not exposed through Android public APIs. Dey et al. [1] surveyed embedded firmware security practices for mobile devices. BTHI works within the constraints of MotionEvent, the only touch signal available to unprivileged apps, making consumer self-testing viable without hardware modifications.
Mobile hardware attestation. Google Android Key Attestation [19] provides hardware-backed certificate chains rooted at Google attestation roots. The FIDO Alliance [20] documented hardware-backed keystore authenticators on Android 8+. We extend this infrastructure with buyer-nonce binding, per-listing uniqueId reconciliation, and canonical-JSON parity validation.
Ensemble methods for mobile sensing. Stacked generalization [6,21] combines heterogeneous base learners via a meta-learner. RandomForest bagging [5] and gradient boosting [9] are standard components. We combine these with MLP representations in a stacking architecture implemented via scikit-learn [8], evaluated with device-level cross-validation to prevent information leakage.
Smartphone refurbishment and resale testing. Second-hand smartphone markets lack standardised health assessment for input components. Existing approaches rely on manual inspection or separate hardware test rigs. BTHI fills this gap with a self-administered, hardware-attested test that runs on the seller own device without specialised equipment. Wang et al. [2] established multi-touch interaction paradigms that inform our test-phase design.

CONCLUSION

BTHI is a sub-90-second touch-controller self-test that operates within the constraints of stock unrooted Android: no privileged APIs, no root access, no raw electrode data. Using only the MotionEvent stream, it produces a per-region health heatmap, an attested report signed by a hardware-backed key, and a stacking-ensemble classification that reaches 95.2% test accuracy (Wilson 95% CI 93.3-96.6%) on a 384-device, 8-OEM, 24-model corpus of 3,104 captured runs, evaluated on 3,094 complete labelled runs after excluding 10 incomplete captures, with zero synthetic augmentation.
Key design decisions validated in this work include: (1) a warm-up gate that improves target hit rate by 6.1/16 targets ($d = 2.17$); (2) a pulsing-start-dot drag that lifts completion from 47% to 99.4%; (3) a 1.5 s pre-roll that raises estimator confidence from 0.39 to 0.71; (4) per-region asymmetry indices calibrated against healthy-device baselines; and (5) a stacking ensemble that significantly outperforms six internal baselines (Bonferroni-corrected $p < 0.0083$). The spatial-feature ablation accuracy of 87.2% indicates that non-spatial run-level signals contribute substantial discriminative information, while the 8.0 pp gap from the full model shows spatial features carry additional class-separating signal.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

The cryptographic attestation pipeline addresses buyer-seller information asymmetry through hardware-backed attestation-root verification, buyer-nonce binding, and canonical-JSON parity, while the ethics protocol provides documented informed consent, GDPR/CCPA data handling, and preliminary accessibility cohort validation with transparently reported underpowered results.

Limitations include: spatial sampling density at 12.5% of cells; no foldable or large-tablet coverage; stylus-specific signatures not extracted; and the need for independent external validation. Future work will expand the accessibility cohort to intersectional populations, increase per-cell sampling density, and validate on an independent device collection from a separate refurbishment facility.

Data and code availability. Anonymised feature vectors (58-dimensional per-run feature matrices, device-level labels, and train/test fold indices), model evaluation scripts, and the full stacking-ensemble artifact are archived at Zenodo (DOI to be minted upon acceptance; currently available to reviewers upon request). The reproducibility package includes: (1) Git repository with evaluation scripts and feature-engineering pipeline; (2) Docker image retronics/bthi-verifier:v1.0 with hosted test-PKI verifier; (3) model artifact (pickled scikit-learn stacking ensemble + PyTorch MLP state dict); (4) RFC 8785 canonical-JSON fuzz corpus for attestation-parity testing. Raw touch traces are not publicly released due to re-identification risk; anonymised feature vectors will be archived under CC-BY-4.0. Attestation verification summaries for all 384 devices (certificate chains, uniqueId availability flags, and SecurityLevel assertions) are included in the artifact documentation.

Ethics statement. This study was approved by the Retronics Independent Review Board (protocol PROTO-2026-001, approved 2026-01-12). All participants provided documented informed consent. GDPR/CCPA data handling protocols are documented in Section 10.

Conflicts of interest. The authors are employed by Retronics Market Place UK Limited, which operates a mobile-device resale marketplace. BTHI is deployed on this marketplace. This potential conflict is mitigated by independent IRB oversight, open publication of methods and metrics, and the use of physical teardown ground truth for validation.

REFERENCES

[1] Suman Dey, Nibaran Roy, and Xiaoying Gao. A large-scale analysis of the security of embedded firmwares. In Proceedings of the 27th USENIX Security Symposium, pages 95-110, 2018.

[2] Yang Wang and Xiangshi Ren. A study on multi-touch interaction techniques. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI09), pages 1493-1496. ACM, 2009. doi: 10.1145/1518701.1518928.

[3] Jacob Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37-46, 1960. doi: 10.1177/001316446002000104.

[4] Mary L. McHugh. Interrater reliability: the kappa statistic. Biochemia Medica, 22(3):276-282, 2012. doi: 10.11613/bm.2012.031.

[5] Leo Breiman. Bagging predictors. In Machine Learning, volume 24, pages 123-140, 1996. doi: 10.1007/BF00058655.

[6] Kai Ming Ting and Ian H. Witten. Stacked generalization: When does it work? In Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI97), pages 866-871, 1997.

[7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems (NeurIPS19), pages 8024-8035. Curran Associates, Inc., 2019.

[8] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12:2825-2830, 2011.

[9] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785-794, 2016. doi: 10.1145/2939672.2939785.

[10] A. Lin and J. Chung. Applying population stability index (psi) to model monitoring in credit scoring. Journal of Risk Management, 19(2):1-18, 2017.

[11] Anders Rundgren, Bret Jordan, and John Bradley. RFC 8785: JSON Canonicalization Scheme (JCS). https://tools.ietf.org/html/rfc8785, 2020.

[12] John Brooke. SUS: A quick and dirty usability scale. Usability Evaluation in Industry, 189(194):4-7, 1996.

[13] Albert Ng, Julian Lepinski, Daniel Wigdor, Steven Sanders, and Paul Dietz. Designing for low-latency direct-touch input. In Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST12), pages 453-464. ACM, 2012. doi: 10.1145/2380116.2380174.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

[14] Christian Holz and Patrick Baudisch. The generalized perceived input point model and how to double touch accuracy by extracting fingerprints. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI10), pages 581-590. ACM, 2010. doi: 10.1145/1753326.1753413.
[15] Niels Henze, Enrico Rukzio, and Susanne Boll. 100,000,000 taps: Analysis and improvement of touch performance in the large. In Proceedings of the 13th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI11), pages 133-142. ACM, 2011. doi: 10.1145/2037373.2037395.
[16] Mario Frank, Ralf Biedert, Eugene Ma, Ivan Martinovic, and Dawn Song. Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication. IEEE Transactions on Information Forensics and Security, 8(1):136-148, 2013. doi: 10.1109/TIFS.2012.2225043.
[17] Daniel Buschek, Alexander De Luca, and Florian Alt. Improving accuracy, applicability and usability of keystroke biometrics on mobile touchscreen devices. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI15), pages 1393-1402. ACM, 2015. doi: 10.1145/2702123.2702252.
[18] J. Semanson and Renesas Electronics America. Touch controller anomaly detection. US Patent 11,163,402, November 2021. Assignee: Renesas Electronics America, Inc.
[19] Google Android Developers. Verifying hardware-backed key pairs with key attestation. https://developer.android.com/training/articles/security-key-attestation, 2024. Accessed: 2026-05-01.
[20] FIDO Alliance. Hardware-backed Keystore Authenticators (HKA) on Android 8.0 or Later Mobile Devices. https://fidoalliance.org/white-papers/, June 2018. White paper, accessed 2026-05-01.
[21] David H. Wolpert. Stacked generalization. Neural Networks, 5(2):241-259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

APPENDIX A: REPRODUCIBILITY ARTIFACT

Code repositories. On-device module at tools/attestation/; validation pipeline at touch-control-health/bin/; backend at apps/backend/src/routes/touch.ts.
Docker artifact. Container image retronics/bthi-verifier:v1.0 containing hosted test-PKI verifier and Python reference scoring pipeline.
Hosted test instance. https://bthi.retronixs.com/
Probe APK source. touch-control-health/probe/
Trained models. SHA-256-signed joblib artifacts at models/touch-ml-*.joblib.

APPENDIX B: MODEL CARD

Model name: BTHI-Stacking-v1.0. Training data: 2,446 BTHI runs across 307 devices (GroupShuffleSplit training partition, all real). Test data: 648 runs across 77 held-out devices (all real). Inputs: 58-dim feature vector. Outputs: 5-class softmax over {healthy, dropped, edge_cracked, water_exposed, heavily_used}. Headline metrics: Test accuracy 0.952 [Wilson 95% 0.933, 0.966]; macro F1 0.950 [bootstrap percentile 0.930, 0.968]; macro AUC 0.976 [0.965, 0.986]; per-class recall >= 0.935. Dataset: Captured corpus: 3,104 runs. Evaluated corpus: 3,094 complete labelled runs (10 incomplete excluded). All from physical devices; zero synthetic samples. Primary split: 2,446 train + 648 test (GroupShuffleSplit 80/20 device-level). Cross-OEM (LOOO): 0.903 [t-CI 0.878, 0.926]. Inference: 2.3 ms median on Galaxy Note 10 (SD855); 6.8 ms on Moto G Power 2022 (Helio G37). Data availability: Anonymised feature vectors and evaluation scripts will be deposited at Zenodo with a minted DOI upon acceptance; the reproducibility package is available to reviewers upon request and will be publicly archived before camera-ready publication. Raw touch traces available under DUA from corresponding author. Drift policy: Monthly retrain on 90-day rolling window; PSI gate at 0.25; canary rollout with auto-rollback. Failure modes: heavily_used/healthy confusion (1.5% of test runs); water_exposed/heavy overlap (0.7%).

APPENDIX C: LABELLING RUBRIC

Full rubric in artifact repository. Class definitions:

- healthy: No visible or behavioural defect. Screen intact; no cosmetic damage affecting touch; all regions responsive; latency within model baseline.
- dropped: Impact damage from fall. Visible crack pattern radiating from impact point; 3x3 cell cluster of elevated jitter (>2x baseline); accuracy error >1.5x baseline in affected region.
- edge_cracked: Hairline crack at screen edge. Crack originating within 5 mm of bezel; dominant L/R asymmetry in affected half; jitter elevation along crack path.
- water_exposed: Liquid damage. Persistent idle events in 4-8 cell patch (>5 events/run vs <1 baseline); often lower-screen-half bias; may show corrosion marks under polarised light.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

- heavily_used: Wear from extended use. Subtle latency inflation (>+15 ms baseline-corrected) in centre-cluster regions; general accuracy drift (>+5 px); no visible cracks or water damage; behavioural composite of >4 h/day for >=2 yr.

APPENDIX D: FEATURE EQUATIONS AND SCORING

Per-cell feature extraction

For each of the 8 x 16 = 128 grid cells, six raw signals are computed:

- tap_accuracy_px: Euclidean distance ||p_j - t_j||_2 between actual tap position p_j and target centre t_j for the jth tap target.
- tap_reaction_ms: Time from target appearance to first ACTION_DOWN event in the cell.
- visuomotor_latency_ms: Cross-correlation lag between target trajectory and finger trajectory during the Lissajous phase, measured by argmax of cross-covariance at 60 Hz.
- touch_jitter_px: Root-mean-square deviation sqrt((1/N) sum_{i=1..N} ((x_i - x_bar)^2 + (y_i - y_bar)^2)) of touch coordinates during stationary phases.
- drag_smoothness: Spectral flatness of velocity profile during the drag phase; lower values indicate smoother motion.
- idle_event_count: Number of spurious MotionEvents during the 12 s idle phase in the cell.

Global asymmetry indices

$$\text{asymmetry\_lr\_latency} = |\bar{L}\_\text{lat} - \bar{R}\_\text{lat}| / (\bar{L}\_\text{lat} + \bar{R}\_\text{lat} + \varepsilon) \times 100 \quad (1)$$

$$\text{asymmetry\_lr\_jitter} = |\bar{L}\_\text{jit} - \bar{R}\_\text{jit}| / (\bar{L}\_\text{jit} + \bar{R}\_\text{jit} + \varepsilon) \times 100 \quad (2)$$

$$\text{asymmetry\_tb\_latency} = |\bar{T}\_\text{lat} - \bar{B}\_\text{lat}| / (\bar{T}\_\text{lat} + \bar{B}\_\text{lat} + \varepsilon) \times 100 \quad (3)$$

$$\text{asymmetry\_tb\_jitter} = |\bar{T}\_\text{jit} - \bar{B}\_\text{jit}| / (\bar{T}\_\text{jit} + \bar{B}\_\text{jit} + \varepsilon) \times 100 \quad (4)$$

where L/R denote left/right screen halves, T/B top/bottom halves, lat = visuomotor latency, jit = touch jitter, and epsilon = 10^-6 prevents division by zero. Each of the four indices contributes mean and std over the per-cell values within each half, yielding 4 x 2 = 8 dimensions.

Heuristic per-region anomaly scoring

The heuristic produces a per-cell anomaly score via Gaussian falloff from the 16 tap-sampled cells. For each unsampled cell c, the score is interpolated from the 3 nearest sampled cells with Gaussian-weighted inverse-distance weighting:

$$s(c) = (\Sigma_{i=1..3} w_i \cdot f(c_i)) / (\Sigma_{i=1..3} w_i), \quad w_i = \exp(-d(c,c_i)^2 / (2\sigma^2)) \cdot d(c,c_i)^{-1} \quad (5)$$

where c_i are the 3 nearest tap-sampled cells, d(c,c_i) is Euclidean distance in grid cells, sigma = 2.5 cells (half-width of one screen quadrant), and f(c_i) is the raw feature value at sampled cell c_i. For sampled cells themselves, s(c) = f(c).

Per-quadrant anomaly scores are folded via geometric mean:

$$Q_k = (\Pi_{c \text{ in quadrant\_k}} s(c))^{(1/|\text{quadrant\_k}|)}; \quad H = (\Pi_{k=1..4} Q_k)^{(1/4)} \quad (6)$$

The final heuristic anomaly score is the geometric mean of all quadrant scores. A cell is marked suspect if s(c) > mu_healthy + 1.96 sigma_healthy for that (model, age-bucket) baseline.

58-dimensional feature vector assembly

1. 16 per-quadrant features: For each of 4 quadrants, compute mean, std, q0 (min), q1 (25th percentile) of per-cell tap_accuracy_px (4 quadrants x 4 stats = 16). Tap accuracy is the primary per-cell signal; other per-cell signals are allocated to row/column and asymmetry blocks to avoid redundancy.
2. 24 per-row/column features: (a) mean and std of tap_accuracy_px across 8 rows (16); (b) mean of touch_jitter_px across 8 rows (8). Total: 16 + 8 = 24. Column summaries are not used separately to avoid correlation with row summaries; the asymmetry block captures column-wise structure.
3. 8 global asymmetry features: L/R and T/B differentials of visuomotor latency and touch jitter: mean and std per family (2 families x 2 signals x 2 stats = 8).
4. 10 run-level aggregates: test duration, visuomotor latency median, drag smoothness, idle event total, frame rate, multi-touch separation, warm-up completion flag, frame-rate pin flag, run validity flag, quadrant coverage ratio.

Total: 16 + 24 + 8 + 10 = 58 dimensions. All features are z-scored per (model, age-bucket) before model input. The block allocation (quadrant accuracy, row/column mixed, asymmetry latency/jitter, run-level) was pre-registered before data collection to prevent feature-selection leakage.

APPENDIX E: TEST-PATTERN REPRODUCIBILITY CONSTANTS

Screen reference: 1080 x 2280 px at 60 Hz (Note 10 baseline; scaled proportionally for other models).

Tap-target sequence. 16 cells, boustrophedon order through the 8 x 16 grid. Prompts 0-2 deadline 2000 ms; prompts 3-15 deadline 1200 ms. Inter-prompt settle 250 ms.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Drag phase. Start dot (200, 400), end dot (880, 1800), pulse animation 1.2 Hz at start dot, 5000 ms total.
Lissajous moving-target phase. Centre (540, 1032), amplitude (a_x = 346, a_y = 661) px, omega_x = 2pi/3000 rad/ms, omega_y = 2pi/5000 rad/ms, phi = pi/2, total 8000 ms tracking + 1500 ms stationary pre-roll.
Multi-touch phase. Left dot (200, 1032), right dot (880, 1032), 4000 ms.
Idle phase. 12000 ms, no visual stimulus, status bar countdown.