International Journal of Safety and Security Engineering
Vol. 16, No. 10, October, 2026, pp. 2247-2258
Journal homepage: http://iieta.org/journals/ijsse

Deep Learning Based Touch and Keystroke Biometric Framework for Continuous Authentication

Marwah Mohameed Khaleel Abdullah[1*], Baban Ahmed Mahmood[2]

[1] Department of Computer Science, College of Computer Science and Information Technology, University of Kirkuk, Kirkuk 36001, Iraq
[2] Department of Information Technology, College of Computer Science and Information Technology, University of Kirkuk, Kirkuk 36001, Iraq

Corresponding Author Email: stcm24007@uokirkuk.edu.iq

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ijsse.161005

Received: 19 August 2026
Revised: 20 September 2026
Accepted: 29 September 2026
Available online: 31 October 2026

Keywords:
behavioral biometrics, continuous authentication, deep learning, keystroke dynamics, multimodal fusion, touch dynamics

ABSTRACT

Continuous user authentication reduces post-login risk by repeatedly verifying identity from behavioral evidence. This study develops a multimodal framework combining touch and keystroke biometrics and evaluates discrimination, probability calibration, security-usability operating points, quality-aware fusion, and naturally missing keystroke evidence. The multilayer perceptron (MLP) serves as the principal reference, alongside a lightweight feature-token transformer, quality-aware gated fusion, and availability-aware routing; a single-step long short-term memory (LSTM) model is retained only as a non-temporal auxiliary reference. Experiments use user-specific binary classification with deterministic eight-fold session-disjoint cross-validation and training-only preprocessing on an 18-feature Fusion representation. At threshold 0.5, MLP achieved 96.24% Accuracy, an area under the receiver operating characteristic curve (AUC) of 0.9356, and an equal error rate (EER) of 11.17%; transformer achieved 95.87%, 0.9361, and 11.49%; and quality-aware fusion achieved false acceptance rates (FAR) of 1.39% but false rejection rates (FRR) of 44.58%. Inner-only Platt calibration reduced mean Brier scores to approximately 0.013. Quality-score perturbations kept learned modality weights near 0.40/0.60, while hard availability routing remained operational when keystroke evidence was missing. Illustrative inner FAR targets of 5% and 1% yielded outer-test FAR ranges of 3.19-3.60% and 0.84-1.00%, with corresponding FRR ranges of 30.00-32.50% and 49.75-53.92%. Results support leakage-resistant, application-dependent control across practical deployment security objectives rather than a universally dominant classifier.

1. INTRODUCTION

Authentication at login does not ensure that the same user remains in control throughout an active session, particularly on mobile or shared devices. Continuous authentication addresses this post-login risk by repeatedly assessing identity during normal use rather than relying solely on point-of-entry credentials.

Behavioral biometrics support passive continuous authentication by characterizing how a user interacts with a device. Touch, keystroke, mouse, gait, voice, and related signals can provide unobtrusive identity evidence [1-6], but their reliability may vary with task, device, context, and behavioral drift. Continuous authentication should therefore be evaluated as both a classification and a security-usability problem.

Touch and keystroke dynamics are particularly suitable for mobile authentication because they can be collected without additional hardware [7]. Touch captures spatial, temporal, pressure, and motion-related behavior, whereas keystroke dynamics captures typing rhythm through hold, flight, and key-transition information. Their complementary evidence motivates multimodal fusion, although performance must be interpreted across discrimination and operating-point metrics rather than accuracy alone [1, 2, 8-10].

Key challenges remain: a single behavioral modality may be noisy or incomplete; user-specific authentication is strongly imbalanced; high Accuracy can coexist with unfavorable false acceptance rates (FAR) or false rejection rates (FRR); and raw model scores need not be well calibrated as probability estimates. These issues motivate fusion, training-only imbalance handling, explicit probability calibration, separate operating-point selection, and evaluation with FAR, FRR, equal error rate (EER), area under the receiver operating characteristic curve (AUC), balanced metrics, and calibration diagnostics.

This study evaluates touch-keystroke continuous authentication using a multilayer perceptron (MLP) principal reference baseline and an auxiliary long short-term memory (LSTM) reference informed by the study [1]. The proposed path uses a lightweight transformer, quality-aware gated fusion, and availability-aware routing for naturally missing behavioral evidence. The auxiliary LSTM is a single-step neural reference only and is not used to support claims about

temporal sequence learning. Random Over Sampler (ROS) is restricted to training data. For the final core MLP, transformer, and quality-aware models, inner session-disjoint out-of-fold predictions are used first for explicit Platt probability calibration and then, separately, for FAR-FRR-balanced operating-point selection without outer-test leakage. Performance is assessed with Accuracy, Balanced Accuracy, AUC, EER, FAR, FRR, true acceptance rate (TAR); true rejection rate (TRR), average classification error rate (ACER), Brier score, and reliability analysis.

The contributions are fivefold. First, a fixed user-specific cohort is evaluated with deterministic session-disjoint cross-validation and training-only preprocessing, with probability calibration and operating-point selection kept methodologically distinct. Second, a compact feature-token transformer with explicit feature identity is evaluated on the 18-feature Fusion representation. Third, quality-aware gating is evaluated using a neutral equal-weight quality-score construction together with two symmetric ±20% coefficient-perturbation sensitivity schemes. Fourth, naturally missing keystroke evidence is handled by explicit availability routing and evaluated through hard-mask and quality-channel ablations. Nested threshold selection is treated as an operating-point decision procedure rather than as probability calibration or a new classifier. Fifth, deployment-oriented threshold policy is evaluated with two illustrative inner-validation FAR constraints (5% and 1%), with achieved performance reported only on untouched outer-test data and interpreted through the resulting false-acceptance versus false-rejection cost trade-off.

2. RELATED WORK

Touch and keystroke dynamics are widely studied for mobile continuous authentication because they provide complementary behavioral evidence during ordinary interaction. Touch captures spatial, temporal, pressure, and motion-related characteristics [2, 8, 11], whereas keystroke dynamics uses hold, flight, and key-transition patterns [9, 10]. Both remain sensitive to session and context variation, behavioral drift, sparse input, device changes, and noise [7, 12, 13] motivating cross-session evaluation with FAR, FRR, EER, and AUC.

Multimodal fusion combines behavioral evidence at feature, score, or decision levels [1, 14-16]. The study [1] is particularly relevant because it evaluates feature-level touch-keystroke fusion using MLP and LSTM models, while the studies [14-16] explore dynamic, multimodal, and resource-aware strategies. These studies show that fusion can improve discrimination or robustness, but gains are metric-dependent; higher Accuracy or stronger impostor rejection may coexist with unfavorable genuine-user rejection when class imbalance and operating thresholds are not explicitly addressed.

Recent studies have applied transformer and attention-based architectures to behavioral authentication [17-20]. They mainly operate on temporal keystroke sequences and use datasets, enrolment conditions, and evaluation protocols different from the activity-level touch-keystroke HMOG setting considered here, so their results are contextual rather than direct numerical baselines. Strict cross-session evidence [21] further shows that strong threshold-independent discrimination does not necessarily yield stable operating thresholds on unseen sessions.

Taken together, the literature leaves four practical evaluation needs: reliable session separation, explicit distinction between probability calibration and FAR-FRR operating-point control, controlled verification of learned fusion mechanisms, and robustness when behavioral evidence becomes unavailable. These needs motivate the session-disjoint protocol, inner-only Platt calibration, nested threshold selection, component ablations, and explicit handling of naturally missing keystroke evidence used in this study.

3. METHODOLOGY

The public HMOG dataset comprises 100 users and 24 planned sessions per user (eight reading, eight writing, and eight map-navigation); the present touch-keystroke analysis uses the eight free-text typing sessions with matched evidence [22]. No new participant data were collected. The raw Touch table contained 11,333,974 events, of which one exact full-row duplicate was removed, leaving 11,333,973 events. The raw keystroke table contained 2,507,367 events; removal of 878,064 exact full-row duplicate excess rows left 1,629,303 unique events. No non-identical behavioral record was removed. Activity-level extraction produced 11,528 Touch rows (10 candidate features) and 2,400 keystroke rows (11 features), yielding 2,400 matched rows with 21 candidate Fusion features; excluding the three globally constant Touch pressure summaries left final touch/keystroke/fusion model inputs of 7/11/18 features.

The experimental direction is informed [1] but is not treated as an exact replication. MLP serves as the principal reference baseline. The single-step LSTM is included only as an auxiliary neural reference for reproducibility and is not treated as a temporal sequence model. The transformer and quality/availability-aware components are proposed extensions. For complete-modality experiments, 50 of the 100 eligible HMOG users were randomly selected without replacement using seed 42, independently of model performance, and frozen for final evaluation. The cohort contains 1,200 activity observations from 400 user-session pairs (24 observations per user across the eight typing sessions), and the same cohort and fold map are used for all core models.

For reproducibility, the exact frozen core user IDs are: 100669, 579284, 220962, 777078, 856401, 657486, 218719, 862649, 201848, 771782, 923862, 219303, 278135, 913228, 539502, 796581, 737973, 803262, 733162, 799296, 865881, 538363, 264325, 395129, 710707, 207969, 918136, 751131, 827212, 257279, 892687, 501973, 526319, 973891, 876011, 405035, 863985, 675397, 815316, 561993, 865501, 240168, 785899, 745224, 540641, 342329, 594887, 396697, 525584, 527796.

The frozen core fold map is deterministic: for every selected user, the eight matched typing sessions are ordered by the numeric session suffix and assigned one complete session per outer fold, so session ranks 1-8 map to outer folds 1-8, respectively. This produces 400 frozen user-session assignments and zero train-test user-session overlap. A post hoc descriptive representativeness audit compared the selected and remaining 50 HMOG users using per-user coverage measures and retained touch/keystroke feature means. The largest absolute standardized mean difference was 0.286 for Touch_Sessions; every audited keystroke measure was smaller. The groups are therefore described as broadly comparable rather than distributionally identical.

Availability-aware robustness is evaluated on a separate frozen 50-user cohort selected only for naturally missing keystroke evidence and sufficient session coverage, independently of model performance. The audited outer join contains 11,528 rows (2,400 complete and 9,128 keystroke-missing); restricting it to this cohort yields 6,293 rows (1,200 complete and 5,093 keystroke-missing). Because Touch remains available in these missing cases and the cohort differs from the core cohort, results are interpreted specifically as missing-keystroke robustness rather than as a direct paired ranking against the core models. Eligibility required (i) at least one naturally missing-modality observation, (ii) at least one complete touch-keystroke observation, and (iii) at least eight distinct sessions so that every selected user could contribute to all eight session-disjoint folds. Eligible users were ranked deterministically by missing-modality sample count (descending), then total row count (descending), then User_ID (ascending), and the first 50 were frozen. No model performance measure was used for this selection. Because this rule intentionally enriches the cohort for natural missingness, the availability-aware cohort is not claimed to represent the overall HMOG user population; its results are interpreted only as robustness evidence under naturally missing keystroke input.

3.1 Proposed framework

The framework comprises event cleaning and feature extraction, matched or availability-aware fusion, user-specific training, explicit probability calibration for the final core models, nested operating-point selection, and evaluation. Core experiments use one deterministic eight-fold session-disjoint map, holding out one complete session per user in each fold. StandardScaler, ROS, distribution-dependent quality normalization, Platt fitting, and threshold selection are restricted to training-side data; quality components are reconstructed on-the-fly from the audited events.

Figure 1 summarizes the two branches. The core branch compares MLP, LSTM, transformer, and quality-aware fusion on the matched cohort. MLP, transformer, and quality-aware Fusion are evaluated at threshold 0.5, after inner-only Platt calibration with separate FAR-FRR-balanced threshold selection, and under illustrative inner FAR <= 5% and FAR <= 1% security policies whose thresholds are applied unchanged to outer-test probabilities. The single-step LSTM remains an auxiliary non-temporal reference. The separate availability-aware branch uses the frozen natural missing-keystroke cohort at threshold 0.5 without Platt calibration or nested policy selection.

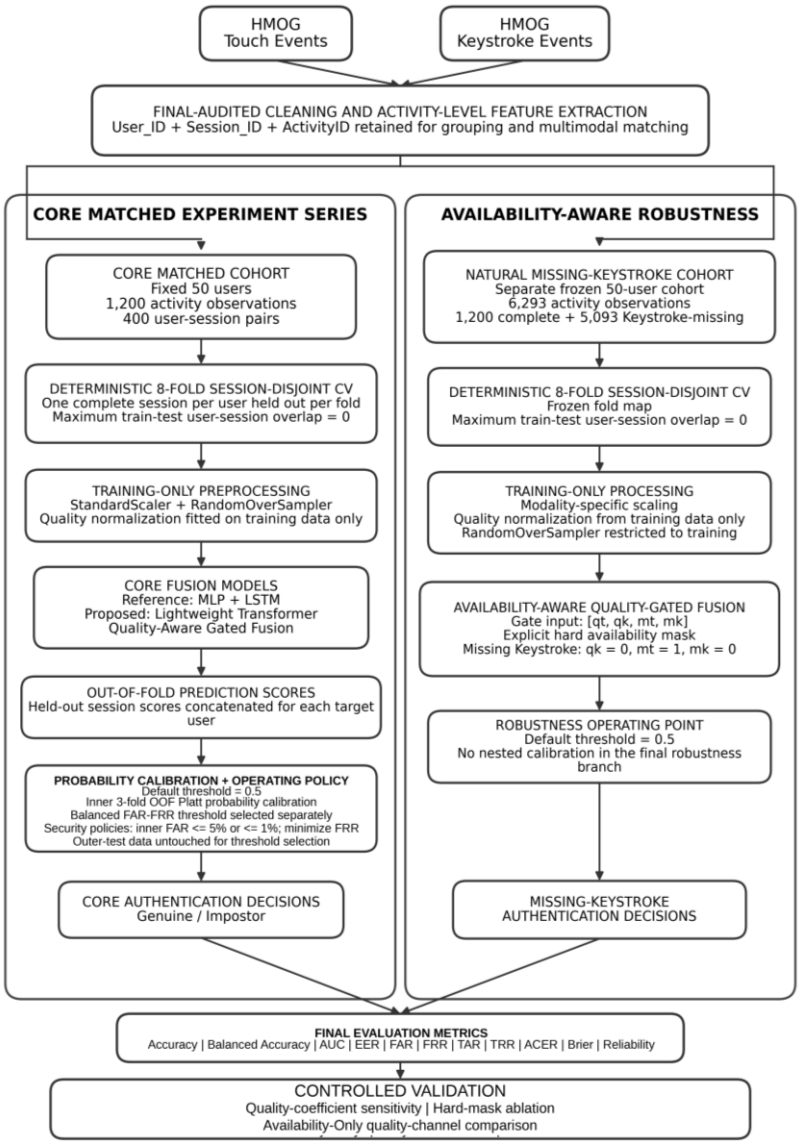


Figure 1. Continuous authentication framework

3.2 Preprocessing and feature engineering

User_ID, Session_ID, and ActivityID are preserved as grouping and multimodal matching keys. Cleaning removes only exact full-row duplicates, converts required fields to valid numeric form, and applies positive-time and finite-value checks before feature derivation. No balancing, scaling, threshold selection, or missing-modality replacement occurs during raw cleaning. The final unimodal touch and keystroke activity-level tables contain no missing or infinite values and no duplicate merge-key combinations.

Raw event logs are converted to authentication features before activity-level aggregation. Touch records provide temporal, spatial, pressure, contact, pointer, and orientation information; keypress records provide PressTime, press type, KeyID, activity, and orientation. Touch aggregation produced 10 candidate features. Pressure_Mean, Pressure_Median, and Pressure_Std were globally constant and were therefore excluded from the model input rather than retained as zero-contribution standardized variables. The resulting representations contain 7 touch features, 11 keystroke features, and 18 matched Fusion features. Importantly, these model inputs are activity-level aggregated vectors: timing-related information is represented by summary features rather than by an ordered event sequence.

Within each user-session-activity group, Touch events are ordered by EventTime and restricted to the same pointer and stroke before differences are computed; Systime is used only as a tie breaker. Non-positive time differences are excluded, and movement distance, velocity, touch intensity, and spatial pressure are derived from valid consecutive events:

$$\Delta t_i = t_i - t_{i-1} \tag{1}$$

$$d_i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \tag{2}$$

$$v_i = \frac{d_i}{\Delta t_i} \tag{3}$$

$$\text{Intensity}_i = \text{Pressure}_i \times v_i \tag{4}$$

$$\text{SpatialPressure}_i = \frac{\text{Pressure}_i}{d_i}, \qquad for\ d_i > 0 \tag{5}$$

The derived Touch quantities are aggregated by User_ID, Session_ID, and ActivityID. Audit of the aggregated table confirmed that Pressure_Mean, Pressure_Median, and Pressure_Std are globally constant and contain no sample-specific information; these three variables are excluded from the model representation. The final 7-feature Touch input therefore consists of Velocity_Mean, Velocity_Max, Velocity_Std, Time_Mean, Time_Median, Intensity_Mean, and Spatial_Pressure_Mean.

Keystroke events are ordered by User_ID, Session_ID, ActivityID, PressTime, Systime, KeyID, and PressType. PressType 0/1 denotes key-down/key-up. In this study, flight time is operationally defined as the press-to-press (down-down) interval between consecutive key-down events, whereas hold time pairs a key-up with the preceding key-down of the same KeyID; only positive intervals are retained. Press_Count, Unique_KeyID, Rhythm_Ratio, and Key_Diversity_Speed (KDS) provide additional rhythm descriptors:

$$\text{Flight}_i = \text{PressTime}_i^{\text{down}} - \text{PressTime}_{i-1}^{\text{down}} \tag{6}$$

$$\text{Hold}_j = \text{PressTime}_j^{\text{up}} - \text{PressTime}_j^{\text{down}} \tag{7}$$

$$\text{Rhythm}_{\text{Ratio}} = \frac{\text{Hold}_{\text{Median}}}{\text{Flight}_{\text{Median}} + 10^{-5}} \tag{8a}$$

$$\text{KDS} = \frac{\text{Unique_KeyID}}{\text{Total_Flight_Time} + 10^{-5}} \tag{8b}$$

The pairing procedure produced 716,933 valid hold intervals, a 91 ms median hold duration, and approximately 99.8% mean group-level hold coverage. The 11 keystroke features are Flight_Mean, Flight_Median, Flight_Std, Hold_Mean, Hold_Median, Hold_Std, Hold_Max, Press_Count, Unique_KeyID, Rhythm_Ratio, and KDS; together with the 7 retained Touch features they form the 18-feature Fusion vector.

3.3 Fusion strategies

Complete fusion uses a one-to-one inner merge of touch and keystroke activity tables on User_ID, Session_ID, and ActivityID, preventing many-to-many duplication. This yields 2,400 matched rows for 100 users; the frozen core cohort contains 1,200 of these rows.

Availability-aware assembly uses an outer join, retaining 2,400 complete and 9,128 naturally keystroke-missing rows, with no naturally missing-Touch counterpart. An unavailable modality is represented by a zero feature vector and availability indicator 0. The frozen availability cohort contains 1,200 complete and 5,093 keystroke-missing rows.

The quality and availability vector for each sample is given by:

$$q = [q_t, q_k, m_t, m_k] \tag{9}$$

where, q_t and q_k are sample-local quality scores and m_t and m_k indicate modality availability. Quality components are reconstructed from audited events; count, duration, distance, and diversity are normalized using training-only 5th/95th percentiles, and stability is computed as 1/(1 + |CV|). The primary formulation uses a neutral equal-weight construction rather than unsupported unequal manual coefficients. Touch quality combines five sample-varying components—event count, duration, travel distance, velocity stability, and completeness—with weight 0.20 each; pressure stability is excluded. Keystroke quality combines event count, duration, key diversity, flight stability, hold stability, and completeness with equal weight 1/6 each. Two symmetric ±20% perturbation schemes around these equal weights are retrained as a sensitivity analysis while keeping the cohort, folds, preprocessing, architecture, and seeds fixed. The quality-construction coefficients are not learned from authentication labels or optimized using test-set performance; adaptive modality weighting is learned separately by the gating network from the resulting quality and availability information. Scores are clipped to [0, 1].

3.4 Reference and proposed models

MLP is used as the principal reference baseline, and an auxiliary LSTM reference is included for reproducibility

relative to the study [1], but neither implementation is treated as an exact replication. In the present implementation, MLP classifier uses one hidden layer of three logistic units, SGD with a constant learning-rate schedule (learning rate 0.3, momentum 0.2), Nesterov momentum disabled, max_iter = 500, and random_state = 42; remaining parameters retain scikit-learn defaults. The auxiliary LSTM uses a compact two-unit recurrent layer, following the reference architecture in the study [1]. The source study reports the LSTM architecture, but it does not specify an equivalent temporal sequence construction for the present activity-level HMOG representation. In our implementation, each aggregated activity vector is reshaped to sequence length 1 and passed to a two-class SoftMax output. The LSTM is trained using Adam (learning rate 0.001) for 10 epochs with batch size 32 and 10% validation, with fold-specific deterministic seeds. A one-step input does not permit the recurrent mechanism to learn temporal dependencies. Accordingly, the LSTM is treated only as a descriptive neural reference for reproducibility and is not interpreted as a temporal event-sequence model or used to support claims about temporal modeling.

The proposed transformer treats each scalar behavioral feature as a token. Let X denote the scalar feature-token input, $E(\cdot)$: the shared dense projection, P: the learned feature-index embedding, $T(\cdot)$: the transformer encoder, and Z: the encoded token representation. The mapping is:

$$Z = T(E(X) + P) \tag{10}$$

The transformer uses two attention heads (key dimension 8), a 32-unit feed-forward layer, dropout 0.10, and a 16-unit ReLU head with two-class SoftMax. For the 18-feature Fusion representation, it contains 2,850 trainable parameters; lightweight therefore refers to compact parameterization, not measured latency, memory, or energy efficiency. Training uses Adam (0.001), sparse categorical cross-entropy, 10 epochs, batch size 32, 10% validation, shuffling, and deterministic target-user/fold seeds.

Quality-aware gated fusion uses separate 16-dimensional touch and keystroke encoders. Each modality-specific encoder consists of a 16-unit ReLU dense layer, layer normalization (epsilon = 10^{-6}), and a second 16-unit ReLU projection. Let $q = [q_t, q_k, m_t, m_k]$ be the quality-availability vector defined in Section 3.3. The gating network $G(\cdot)$ comprises a Dense(8, ReLU) hidden layer followed by a two-unit output projection; SoftMax converts the resulting two logits into learned touch and keystroke weights w_t and w_k:

$$[w_t, w_k] = \mathrm{softmax}(G(q)) \tag{11}$$

To enforce availability, let $j \in \{t, k\}$, where t denotes Touch and k denotes keystroke. Each learned weight is masked by the binary indicator m_j and normalized as:

$$\alpha_j = \frac{w_j m_j}{w_t m_t + w_k m_k + \varepsilon}, j \in \{t, k\} \tag{12}$$

where, α_j is the effective normalized weight and $\varepsilon = 10^{-7}$ is used specifically for numerical stability in modality-weight normalization. Let e_t and e_k denote the 16-dimensional touch and keystroke encoder outputs, and let $\|$ denote concatenation. The final fusion vector is:

$$F = [\alpha_t e_t \| \alpha_k e_k \| q_t \| q_k \| m_t \| m_k] \tag{13}$$

Because e_t and e_k are each 16-dimensional and the quality-availability vector contributes four additional values, F contains 36 scalar components. The resulting 36-dimensional fusion vector is reshaped into feature tokens and processed by the same transformer block design described above, including feature-index embedding, multi-head self-attention, the feed-forward block, global average pooling, and the two-class classification head.

In the core matched cohort $m_t = m_k = 1$, so quality-aware fusion evaluates quality-conditioned weighting with both modalities present. The primary quality descriptors use neutral equal component weights with pressure stability excluded. Sensitivity is assessed by retraining the same quality-aware architecture under two symmetric ±20% coefficient-perturbation schemes, with the cohort, fold map, preprocessing, architecture, and seeds held fixed. The gate itself remains learned and sample-adaptive; the sensitivity experiment therefore tests dependence on the manually specified quality-score construction rather than replacing the learned modality gate.

Availability-aware fusion adds explicit hard routing to the same quality/gating mechanism. When keystroke is naturally absent, its feature vector and q_k are set to zero, $m_t = 1$ and $m_k = 0$. In the implemented availability-aware model, the raw modality vectors are additionally multiplied by their availability indicators before entering the corresponding modality encoders, preventing unavailable evidence from propagating through the encoder path. Before the normalization in Eq. (12), the learned weights are masked as:

$$\tilde{w} = [w_t m_t, w_k m_k] \tag{14}$$

With missing keystroke, hard routing forces its effective weight to zero and renormalizes Touch toward one. The no-hard-availability-mask ablation keeps data, quality/availability inputs, scaling, ROS, architecture, folds, and seeds fixed while removing only the hard weight mask, thereby isolating explicit routing.

The availability-only ablation isolates sample-specific quality evidence while preserving the frozen availability cohort, fold map, preprocessing, architecture, seeds, true availability indicators, and hard mask. Only the gate input changes: Full availability-aware uses $[q_t, q_k, m_t, m_k]$, whereas availability-only uses $[0, 0, m_t, m_k]$. Both are evaluated at threshold 0.5 without Platt probability calibration or nested threshold selection.

3.5 Training protocol, probability calibration, and threshold selection

For each target user, that user's observations are treated as genuine, whereas observations from the other 49 users are treated as impostors. StandardScaler is fitted only on the training partition, after which Random Over Sampler(random_state = 42) balances the training classes to 1:1. Test data are neither resampled nor used to estimate scaling parameters.

Core evaluation uses deterministic eight-fold session-disjoint cross-validation rather than row-level stratification. Each fold holds out one complete session from every user, yielding 3 genuine and 147 impostor activity observations (1:49) for each target user. This natural test prevalence is preserved, and train-test user-session overlap is required to be zero. The separate availability cohort uses its own fixed eight-

fold session-disjoint fold map under the same zero-overlap requirement.

The raw model decision threshold is 0.5. For the final MLP, transformer, and quality-aware fusion models, explicit probability calibration is performed entirely inside each outer-training partition. Three inner session-disjoint folds generate out-of-fold raw scores on the natural 1:49 class distribution. A sigmoid (Platt) calibrator is then fitted to those inner out-of-fold scores and labels and applied unchanged to the corresponding untouched outer-test raw scores. After probability calibration, the operating threshold is selected separately on the calibrated inner predictions by minimizing the absolute difference between FAR and FRR:

$$\tau^* = \arg\min_{\tau} |\mathrm{FAR}(\tau) - \mathrm{FRR}(\tau)| \tag{15}$$

If multiple thresholds produce the same minimum difference, ties are resolved deterministically by lexicographically minimizing the following tuple, in the stated order:

$$B(\tau) = (\mathrm{ACER}(\tau), \mathrm{FAR}(\tau), |\tau - 0.5|) \tag{16}$$

The selected threshold for each target-user/fold pair is then applied only to the corresponding untouched outer-test calibrated probabilities. Probability calibration and threshold selection serve different purposes: Platt scaling transforms the score into a calibrated probability estimate, whereas threshold selection changes the binary operating rule. Because the fitted Platt map is monotonic, discrimination metrics such as AUC and EER are evaluated from the original score ranking and are not claimed to improve after calibration. Calibration quality is evaluated separately using Brier score, log loss, and reliability curves. The EER threshold is computed post hoc from outer-test score rankings for diagnostic reporting, whereas the FAR-FRR-balanced operating threshold is selected exclusively from inner data.

To evaluate deployment-oriented security policy separately from the balanced analytical point, two additional illustrative operating constraints are applied to the same calibrated inner out-of-fold predictions. For an inner FAR cap $\tau \in \{0.05, 0.01\}$, the selected security threshold minimizes FRR subject to FAR $\leq \tau$; deterministic ties are resolved by lower FAR and then the higher threshold. The selected threshold is then applied unchanged to the corresponding untouched outer-test calibrated probabilities. The 5% and 1% values are illustrative policy targets rather than universal biometric standards, and outer-test FAR is reported as the achieved value rather than assumed to satisfy the inner constraint.

Reproducibility is supported by fixed user selection, deterministic fold assignments, and predefined random seeds. MLP/ROS uses random_state = 42, whereas neural models use deterministic target-user/fold seeds defined as $s = (42 + 31u + 997f) \bmod (2^{31} - 1)$, where u denotes the target-user identifier and f denotes the fold number. The exact 50-user core membership and the 400 user-session outer-fold assignments are also frozen; the corresponding audit verified a maximum train-test user-session overlap of zero.

3.6 Evaluation metrics

Metrics are Accuracy, Balanced Accuracy, AUC, EER, TAR, TRR, FAR, FRR, ACER, Brier score, and log loss. Balanced Accuracy averages genuine sensitivity and impostor specificity; FAR/FRR measure impostor acceptance and genuine rejection; TAR = 1 - FRR, TRR = 1 - FAR, and ACER = (FAR + FRR)/2. AUC and EER characterize score-level discrimination independently of a fixed decision threshold. Brier score and log loss quantify probability-calibration quality, and reliability curves compare mean predicted genuine probability with the observed genuine fraction using quantile-binned out-of-fold predictions.

For each target user, scores from all eight outer test folds are concatenated before user-level metrics are calculated; final values are means across the 50 users. This preserves complete session-disjoint out-of-fold evaluation at the user level.

For security-constrained operating-policy analysis, FAR, FRR, TAR, TRR, Balanced Accuracy, and Accuracy are computed only after applying the inner-selected threshold to the untouched outer-test predictions. Because the FAR constraint is imposed on inner validation data, the reported outer-test FAR is an achieved generalization result and is not forced to equal the nominal 5% or 1% target.

Availability-aware metrics are reported separately for Overall, complete-modality, and keystroke-missing scopes. Because its cohort differs from the core matched cohort, these absolute results are interpreted as robustness evidence rather than as a direct paired ranking of core models.

Full availability-aware versus availability-only is compared at the target-user level using two-sided paired Wilcoxon signed-rank tests for seven inferential metrics (Accuracy, Balanced Accuracy, AUC, EER, FAR, FRR, and ACER), with a global Holm correction across 21 tests (three scopes by seven metrics). TAR and TRR are reported descriptively as complements of FRR and FAR and are therefore not tested separately. Paired mean differences are accompanied by reproducible 20,000-resample percentile-bootstrap 95% confidence intervals and matched-pairs rank-biserial correlations, direction-adjusted so positive values favor the Full model. Bootstrap intervals are descriptive; Holm-adjusted Wilcoxon inference is primary. Balanced Accuracy and ACER are mathematical complements; both are reported for interpretability, but their inferential results are not interpreted as independent evidence.

4. RESULTS AND DISCUSSION

This section reports core model performance, threshold-dependent operating behavior, and controlled ablations under the frozen session-disjoint protocol defined in Section 3.

4.1 Experimental setup

Core results use the fixed 50-user matched cohort (1,200 observations), whereas availability-aware results use the separate 50-user natural-missing cohort (6,293 observations). Core test folds contain 3 genuine and 147 impostor observations per target user; therefore, Accuracy is interpreted jointly with Balanced Accuracy, FAR, FRR, AUC, and EER.

All final evaluations use deterministic eight-fold session-disjoint splits with training-only scaling and ROS and zero train-test session overlap. The raw default decision threshold is 0.5. For MLP, transformer, and quality-aware fusion, explicit Platt calibration is fitted on three-fold inner session-disjoint out-of-fold predictions and the FAR-FRR-balanced operating threshold is then selected separately from the calibrated inner probabilities. The same leakage-free inner

predictions are additionally used for the illustrative FAR ≤ 5% and FAR ≤ 1% security-policy thresholds, which are evaluated only on the corresponding untouched outer-test folds.

4.2 Overall performance of fusion models

Table 1 summarizes the performance of the core Fusion models at the raw default operating point and, for the final core MLP, transformer, and quality-aware models, after explicit Platt probability calibration followed by nested FAR-FRR-balanced threshold selection.

At the default threshold of 0.5, the quality-aware fusion model achieved the highest Accuracy (97.74%) and the lowest FAR (1.39%), but its Balanced Accuracy was 77.01% and its FRR was 44.58%. This illustrates how raw Accuracy can appear favorable under the 1:49 class imbalance despite a high genuine-rejection rate. MLP achieved 96.24% Accuracy with FAR 3.22%, while the transformer achieved 95.87% Accuracy with FAR 3.64%. Thus, no core model dominates all default-threshold criteria.

At the score level on the 18-feature representation, the transformer produced a slightly higher AUC than MLP (0.9361 vs 0.9356), whereas MLP produced a slightly lower EER (11.17% vs 11.49%). The quality-aware model produced AUC 0.9157 and EER 14.15%. These descriptive differences do not establish statistical equivalence or architectural superiority. The auxiliary LSTM is reported for descriptive reproducibility only and is not used for cross-model inference or interpreted as evidence of temporal modeling. Final probability-calibration and threshold-dependent results are reported separately in Section 4.3 on the same 18-feature representation.

Table 1. Core 50-user matched fusion performance at the raw default operating point and final calibrated nested balanced operating point

Model / Operating Point	Accuracy	Balanced Accuracy	AUC	EER	FAR	FRR	TAR	TRR
MLP + ROS (Reference) - Default 0.5	96.24%	83.35%	0.9356	11.17%	3.22%	30.08%	69.92%	96.78%
MLP + ROS (Reference) - Platt calibrated + nested balanced threshold	89.61%	88.70%	0.9356	11.17%	10.36%	12.25%	87.75%	89.64%
LSTM + ROS (Auxiliary; non-temporal) - Default 0.5	87.37%	85.92%	0.9130	13.71%	12.57%	15.58%	84.42%	87.43%
LSTM + ROS (Auxiliary; non-temporal) - Nested threshold (descriptive only)	86.78%	85.66%	0.9130	13.71%	13.17%	15.50%	84.50%	86.83%
Lightweight transformer (proposed) - Default 0.5	95.87%	84.14%	0.9361	11.49%	3.64%	28.08%	71.92%	96.36%
Lightweight transformer (proposed) - Platt calibrated + nested balanced threshold	90.07%	88.32%	0.9361	11.49%	9.86%	13.50%	86.50%	90.14%
Quality-aware gated (proposed) - Default 0.5	97.74%	77.01%	0.9157	14.15%	1.39%	44.58%	55.42%	98.61%
Quality-aware gated (proposed) - Platt calibrated + nested balanced threshold	88.55%	84.97%	0.9157	14.15%	11.30%	18.75%	81.25%	88.70%

Note: The auxiliary LSTM is included only as a descriptive non-temporal reference. Multilayer perceptron (MLP); Random Over Sampler (ROS); long short-term memory (LSTM); area under the receiver operating characteristic curve (AUC); equal error rate (EER); false acceptance rates (FAR); false rejection rates (FRR); true acceptance rate (TAR); true rejection rate (TRR).

4.3 Security-usability trade-off, probability calibration, and threshold selection

Explicit sigmoid (Platt) calibration materially improved the probability estimates of all three final core models. Mean Brier score decreased from 0.029907 to 0.013206 for MLP, from 0.031210 to 0.013072 for transformer, and from 0.019566 to 0.013404 for quality-aware fusion. Mean log loss likewise decreased from 0.114662 to 0.053305 for MLP, from 0.109846 to 0.052822 for transformer, and from 0.101282 to 0.057540 for quality-aware fusion. The improvement in Brier score occurred for 50/50 MLP users, 49/50 transformer users, and 48/50 quality-aware users; paired Wilcoxon tests on user-level Brier scores were significant for all three models ($p < 0.001$). Figure 2 compares the reliability curves for raw and inner-only Platt-calibrated predictions across (a) MLP, (b) the lightweight transformer, and (c) quality-aware gated fusion. The dashed diagonal represents perfect calibration, and the curves are computed from final out-of-fold predictions using quantile bins.

After calibration, the median FAR-FRR-balanced thresholds were 0.0214 for MLP, 0.0199 for transformer, and 0.0128 for quality-aware fusion, with interquartile ranges 0.0178–0.0267, 0.0166–0.0244, and 0.0098–0.0162, respectively. These values should not be interpreted against a universal 0.5 decision rule. Each user-specific outer test fold contains 3 genuine and 147 impostor observations, preserving an approximately 2% genuine prevalence, and the selected threshold is optimized specifically to balance FAR and FRR rather than to implement a generic equal-cost posterior decision at 0.5. The persistence of low operating thresholds after explicit calibration therefore does not, by itself, indicate failed probability calibration.

At the final calibrated nested balanced operating point, MLP achieved 89.61% Accuracy, 88.70% Balanced Accuracy, FAR 10.36%, FRR 12.25%, TAR 87.75%, and TRR 89.64%. Transformer achieved 90.07% Accuracy, 88.32% Balanced Accuracy, FAR 9.86%, FRR 13.50%, TAR 86.50%, and TRR 90.14%. quality-aware fusion achieved 88.55% Accuracy, 84.97% Balanced Accuracy, FAR 11.30%, FRR 18.75%, TAR 81.25%, and TRR 88.70%. Because Platt scaling is monotonic, AUC/EER remain the discrimination values of the corresponding raw score rankings: 0.9356/11.17% for MLP, 0.9361/11.49% for transformer, and 0.9157/14.15% for quality-aware fusion.

All probability-calibration fitting and threshold selection were restricted to outer-training data. The audit reported zero outer train-test session overlap, zero inner train-validation session overlap, and zero inner-to-outer-test session intrusion. Thus, the calibration and operating-point results are fully out-of-fold with respect to each reported outer test partition.

The balanced FAR-FRR operating point is not treated as a universal deployment threshold because its outer-test FAR remains approximately 9.86-11.30% across the three final core

models. We therefore evaluated stricter security-oriented policies by selecting, within calibrated inner out-of-fold predictions only, the threshold that minimized FRR while satisfying an illustrative FAR cap of 5% or 1%.

The outer-test FAR and FRR values achieved under these illustrative, non-universal policy settings are reported in Table 2.

Under the inner FAR ≤ 5% policy, achieved outer-test FAR decreased to 3.60% for MLP, 3.19% for transformer, and 3.27% for quality-aware fusion, while FRR increased to 30.00%, 30.83%, and 32.50%, respectively (TAR 70.00%, 69.17%, and 67.50%). Under the stricter inner FAR ≤ 1% policy, achieved outer-test FAR was 0.84%, 1.00%, and 0.84%, respectively, but FRR increased to 53.92%, 49.75%, and 52.58% (TAR 46.08%, 50.25%, and 47.42%). These results demonstrate a direct security-usability cost trade-off: lower impostor acceptance requires substantially greater genuine-user rejection.

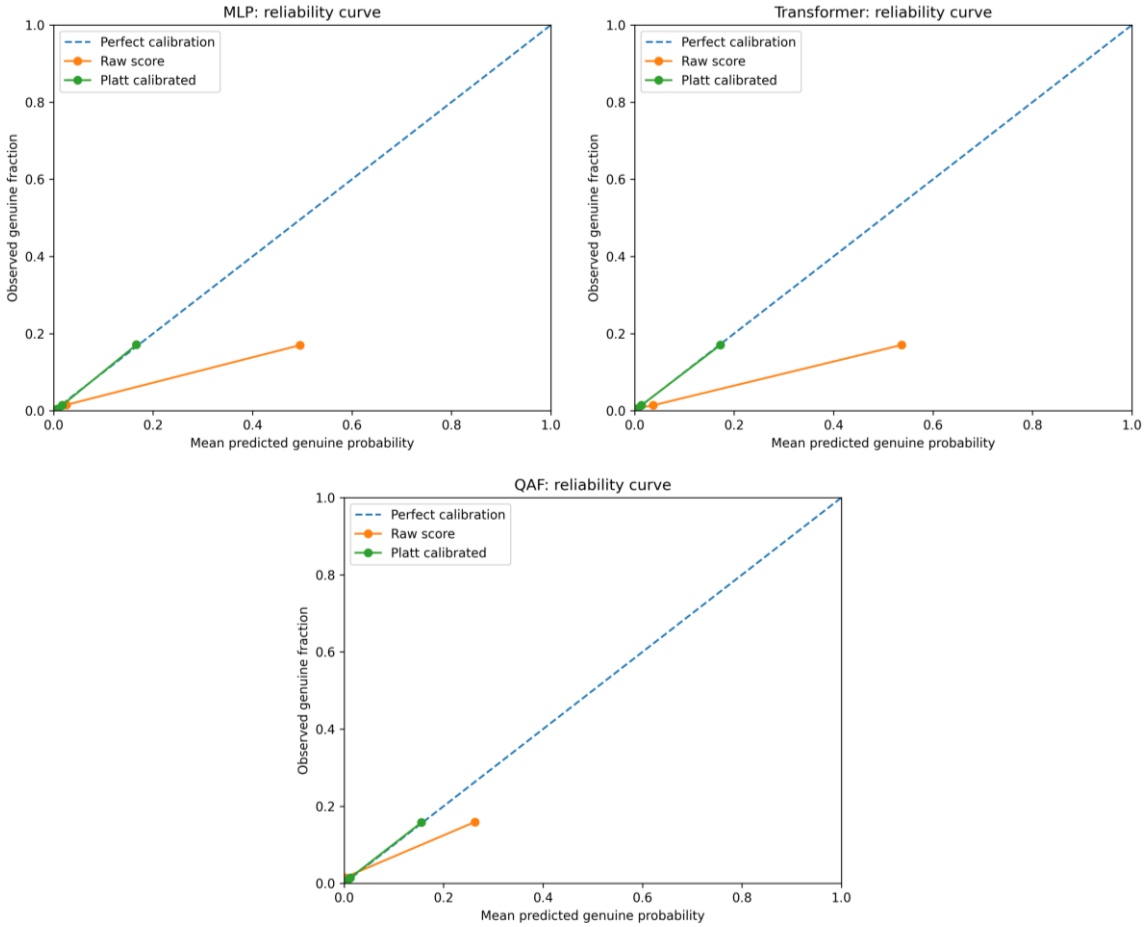


Figure 2. Reliability curves before and after Platt calibration

Table 2. Security-constrained operating-policy analysis

Model	Balanced FAR / FRR	Inner FAR ≤ 5% Outer FAR / FRR	Inner FAR ≤ 1% Outer FAR / FRR
MLP	10.36% / 12.25%	3.60% / 30.00%	0.84% / 53.92%
Transformer	9.86% / 13.50%	3.19% / 30.83%	1.00% / 49.75%
Quality-aware	11.30% / 18.75%	3.27% / 32.50%	0.84% / 52.58%

Note: multilayer perceptron (MLP); false acceptance rates (FAR); false rejection rates (FRR).

4.4 Controlled validation of proposed components

Controlled comparisons used to assess the proposed components are summarized in Table 3.

The complete-modality Fusion models were evaluated on the 18-feature representation under the frozen 50-user cohort and deterministic eight-fold session-disjoint protocol. At threshold 0.5, MLP achieved 96.24% Accuracy, 83.35% Balanced Accuracy, AUC 0.9356, EER 11.17%, FAR 3.22%, and FRR 30.08%; the transformer achieved 95.87%, 84.14%, 0.9361, 11.49%, 3.64%, and 28.08%, respectively. The quality-aware model with equal no-pressure quality coefficients achieved 97.74% Accuracy, 77.01% Balanced Accuracy, AUC 0.9157, EER 14.15%, FAR 1.39%, and FRR 44.58%. Under the two symmetric ±20% coefficient perturbations, AUC was 0.9122 and 0.9128, EER was 14.43% and 14.56%, FAR remained approximately 1.40%, and FRR was 44.58% and 44.83%. Mean learned modality weights were also stable: touch/keystroke = 0.3975/0.6025 for equal weighting, 0.3983/0.6017 for Perturb-A, and 0.3977/0.6023 for Perturb-B. These findings are interpreted as empirical stability within the tested ±20% range, not as proof of coefficient invariance or universal robustness.

Table 4 further evaluates whether sample-specific quality adds value when availability handling is fixed by comparing Full availability-aware with availability-only under naturally missing keystroke.

Table 3. Controlled evidence connecting identified limitations to the proposed components

Research Gap / Component	Controlled Comparison	Specific Experimental Evidence
Session-dependent evaluation	Frozen 8-fold session-disjoint validation	One complete session per user is held out in each core test fold; train-test user-session overlap = 0.
Quality-score coefficient sensitivity	Equal_No Pressure vs two symmetric ± 20% coefficient perturbations	Equal weighting: AUC 0.9157, EER 14.15%, FAR 1.39%, FRR 44.58%, mean Touch/Key weights 0.3975/0.6025. Perturb-A/B: AUC 0.9122/0.9128; EER 14.43%/14.56%; FAR ≈ 1.40%/1.40%; FRR 44.58%/44.83%; learned weights remained close to 0.40/0.60.
Natural keystroke unavailability	Availability-aware: complete-modality vs keystroke-missing within the same cohort	Complete-modality: Accuracy 96.87%, AUC 0.8899, EER 17.57%, FAR 2.26%, FRR 45.67%. Keystroke-missing: Accuracy 82.55%, AUC 0.7292, EER 32.44%, FAR 16.74%, FRR 52.63%.
Explicit hard availability routing	Hard mask vs no-hard-availability-mask, keystroke-missing	FRR 52.63% vs 56.19% (Holm p = 0.0031); FAR 16.74% vs 15.13% (p = 0.000016); Accuracy 82.55% vs 84.06% (p = 0.000059). Without hard masking, absent-Key weight: mean 23.49%, max 70.50%; with hard masking: 0.
Quality evidence under natural missingness	Full availability-aware vs availability-only, keystroke-missing	Full model directionally improved balanced accuracy, AUC, EER, FRR, and ACER; no metric remained significant after global Holm correction (all adjusted p >= 0.923). Paired effects and 95% bootstrap CIs are reported in Table 4.

Table 4. Paired quality-evidence ablation under naturally missing keystroke (n = 50 users)

Metric	Full	Availability-Only	Delta (Full-Only)	95% CI	r_rb*	Holm p
Accuracy	82.55%	82.79%	-0.24 pp	[-0.73, +0.27] pp	-0.189	1.000
Balanced Accuracy	65.32%	64.62%	+0.70 pp	[-0.02, +1.41] pp	+0.327	0.923
AUC	0.7292	0.7253	+0.0039	[-0.0007, +0.0084]	+0.260	1.000
EER	32.44%	32.91%	-0.47 pp	[-1.08, +0.12] pp	+0.224	1.000
FAR	16.74%	16.46%	+0.28 pp	[-0.26, +0.81] pp	-0.190	1.000
FRR	52.63%	54.30%	-1.67 pp	[-3.34, -0.02] pp	+0.299	1.000
ACER	34.68%	35.38%	-0.70 pp	[-1.41, +0.00] pp	+0.327	0.923

Note: * Direction-adjusted matched-pairs rank-biserial correlation; positive values favor the full availability-aware model. Delta and 95% confidence interval (CI) are on the original metric scale (percentage points for rate metrics). Confidence intervals are descriptive paired-user percentile-bootstrap intervals (20,000 resamples). Formal inference uses two-sided paired Wilcoxon tests with global Holm correction across 21 scope-metric tests. No comparison in this ablation remained significant after Holm correction. Balanced Accuracy and ACER are complements.

Under keystroke-missing, the full model directionally improved balanced accuracy (65.32% vs 64.62%), AUC (0.7292 vs 0.7253), EER (32.44% vs 32.91%), FRR (52.63% vs 54.30%), and ACER (34.68% vs 35.38%), while availability-only had slightly higher accuracy and lower FAR. No difference remained significant after global Holm correction (adjusted $p \geq 0.923$); the result is therefore interpreted as modest directional, not statistically conclusive, evidence for the quality channels.

4.5 Discussion

No single core configuration dominates all default-threshold or score-level metrics. Transformer has the slightly higher AUC (0.9361 vs 0.9356), whereas MLP has the slightly lower EER (11.17% vs 11.49%). Quality-aware has the highest default Accuracy and lowest FAR but also a high FRR. After explicit Platt calibration and nested FAR-FRR-balanced threshold selection, MLP and transformer produced similar balanced accuracy (88.70% and 88.32%), while quality-aware reached 84.97%. These operating points substantially reduce genuine rejection relative to the raw default threshold but increase false acceptance, so they should be interpreted as balanced analytical operating points rather than universal deployment thresholds. The proposed transformer is therefore presented as a compact competing architecture rather than as universally superior to the MLP reference.

The security-policy analysis makes the deployment consequence explicit. The FAR-FRR-balanced thresholds reduce genuine rejection but leave false acceptance near 10-11%, so they are analytical trade-off points rather than default security settings. When false acceptance carries a higher application cost, a stricter inner FAR policy can be selected: the illustrative 5% target generalized to 3.19–3.60% outer FAR with FRR near 30–33%, whereas the 1% target generalized to 0.84–1.00% outer FAR but raised FRR to approximately 50–54%. Accordingly, deployment threshold selection should be based on application-specific security and usability costs rather than on a single universal operating point; this study does not estimate a monetary cost function or prescribe the 1% or 5% targets as standards.

Direct comparison with recent keystroke transformer studies [17-20] should remain cautious because their temporal-sequence settings, datasets, and enrolment conditions differ from the present 18-feature touch–keystroke feature-level fusion representation derived from activity-level HMOG observations. In addition, the study [21] shows that threshold-free discrimination can transfer better across sessions than operating thresholds. The present contribution is therefore controlled internal and component-level evidence under a fixed session-disjoint protocol, not a claim of universal state-of-the-art accuracy. In the present work, the core classifiers operate on activity-level aggregated feature vectors rather than ordered raw-event sequences; therefore, no claim of temporal-sequence superiority is made for either the auxiliary LSTM or the proposed transformer.

The quality-aware analysis separates fixed quality-score construction from learned modality gating. With pressure stability excluded, the primary touch and keystroke quality descriptors use neutral equal component weights, while the gating network continues to learn sample-specific touch/keystroke modality weights. Across the two symmetric ±20% coefficient perturbations, changes in AUC, EER, FAR, FRR, and mean learned modality weights were modest; this is

interpreted only as empirical stability within the tested perturbation range, not as coefficient invariance or theoretical optimality. The auxiliary LSTM is not used to draw conclusions about temporal dependency learning.

Availability-aware fusion remains operational when keystroke is missing but with substantial discrimination loss: Accuracy/AUC/EER change from 96.87%/0.8899/17.57% under Complete-Modality to 82.55%/0.7292/32.44% under keystroke-missing. These results are robustness evidence, not a standalone Touch benchmark or direct comparison with the core cohort. Hard routing forces the absent keystroke weight to zero; without it, the learned absent-Key weight averaged 23.49% and reached 70.50%. The hard mask reduced FRR from 56.19% to 52.63% (Holm-adjusted $p = 0.0031$) but increased FAR and lowered raw Accuracy, confirming a security-usability trade-off.

When the hard mask and availability indicators are held fixed, removing q_t and q_k produces only modest changes. The Full availability-aware model directionally improves balanced accuracy, AUC, EER, FRR, and ACER, but no comparison survives the conservative global Holm correction. Thus, sample-specific quality evidence appears useful for several missing-keystroke metrics, but the present paired analysis does not establish statistically significant across-metric superiority.

Overall, the evidence supports leakage-resistant session-disjoint validation, compact feature-token attention on the 18-feature representation, explicit probability calibration with improved Brier scores, separate nested operating-point selection, quality-conditioned fusion with bounded coefficient-sensitivity analysis, and explicit routing under natural keystroke unavailability. The findings are deliberately metric-specific and do not support a claim that one classifier dominates all references or operating objectives. It further shows that balanced and security-constrained thresholds answer different deployment objectives and should be selected according to the relative cost of false acceptance and false rejection.

5. CONCLUSIONS

This study evaluated feature-level fusion of HMOG touch and keystroke behavioral features for continuous authentication using user-specific classification and deterministic eight-fold session-disjoint cross-validation. MLP served as the principal reference baseline; the auxiliary single-step LSTM was included only as a non-temporal neural reference for reproducibility. The proposed path uses an 18-feature fusion representation that excludes the three globally constant touch pressure variables, together with a 2,850-parameter transformer, quality-aware gated fusion, explicit Platt probability calibration, separate nested operating-point selection, and availability-aware routing. Scaling, class balancing, distribution-dependent quality normalization, probability-calibrator fitting, and operating-threshold selection are restricted to training-side data.

In the core cohort at raw threshold 0.5, MLP achieved 96.24% Accuracy, AUC 0.9356, and EER 11.17%, while transformer achieved 95.87%, 0.9361, and 11.49%, respectively. Quality-aware achieved 97.74% default Accuracy and 1.39% FAR but 44.58% FRR, underscoring the need to interpret imbalanced Accuracy together with biometric error rates. Explicit Platt calibration reduced the mean Brier score to 0.0132, 0.0131, and 0.0134 for MLP, transformer, and quality-aware fusion, respectively. The subsequent balanced operating points yielded Balanced Accuracy of 88.70%, 88.32%, and 84.97%, with FAR/FRR of 10.36%/12.25%, 9.86%/13.50%, and 11.30%/18.75%, respectively. These are balanced analytical operating points, not universal security deployment thresholds.

Deployment-oriented threshold analysis confirmed that the balanced operating points are not universal security settings. Using inner-only security constraints, the illustrative FAR $\leq$ 5% policy achieved outer-test FAR of 3.60%, 3.19%, and 3.27% for MLP, transformer, and quality-aware fusion, with FRR of 30.00%, 30.83%, and 32.50%. The stricter FAR $\leq$ 1% policy achieved outer-test FAR of 0.84%, 1.00%, and 0.84%, but FRR increased to 53.92%, 49.75%, and 52.58%. Thus, stronger impostor rejection is achievable at a substantial usability cost, and the deployment threshold should be chosen according to application-specific risk rather than a universal FAR-FRR balance.

The quality-aware formulation remained empirically stable under the tested ±20% perturbations of its neutral quality-score coefficients, with learned touch/keystroke weights remaining close to 0.40/0.60. Under missing keystroke, hard availability routing reduced FRR relative to the no-hard-mask variant but increased FAR, while the quality-channel ablation produced directional rather than Holm-significant gains. These controlled findings are interpreted within their tested conditions and are not generalized beyond the evaluated coefficient range or missing-modality cohort.

The results support explicit treatment of session separation, operating thresholds, quality, and modality availability in behavioral continuous authentication, while showing that robustness costs remain when keystroke evidence is absent. Future work should extend validation to broader user populations and real-time deployment settings.

REFERENCES

[1] Stylios, I., Chatzis, S., Thanou, O., Kokolakis, S. (2023). Continuous authentication with feature-level fusion of touch gestures and keystroke dynamics to solve security and usability issues. Computers & Security, 132: 103363. https://doi.org/10.1016/j.cose.2023.103363

[2] Gattulli, V., Impedovo, D., Pirlo, G., Volpe, F. (2023). Touch events and human activities for continuous authentication via smartphone. Scientific Reports, 13(1): 10515. https://doi.org/10.1038/s41598-023-36780-3

[3] Li, P., Chen, F., Pan, L., Hoang, T., Zhu, Y., Yang, L. (2025). LightLiveAuth: A lightweight continuous authentication model for virtual reality. IoT, 6(3): 50. https://doi.org/10.3390/iot6030050

[4] Chen, H., Xu, Y., Feng, Y., et al. (2026). DynamicLip: Shape-independent continuous authentication via lip articulator dynamics. IEEE Transactions on Mobile Computing, 25(3): 4063-4076. https://doi.org/10.1109/TMC.2025.3624155

[5] Cao, H., Jiang, H., Liu, D., et al. (2023). LiveProbe: Exploring continuous voice liveness detection via phonemic energy response patterns. IEEE Internet of Things Journal, 10(8): 7215-7228. https://doi.org/10.1109/JIOT.2022.3228819

[6] Durmaz, H., Keskinöz, M. (2023). On the improvements of mouse dynamics based continuous user

authentication. In 2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI), Bellevue, WA, USA, pp. 13-19. https://doi.org/10.1109/IRI58017.2023.00010

[7] Baig, A.F., Eskeland, S., Yang, B. (2023). Privacy-preserving continuous authentication using behavioral biometrics. International Journal of Information Security, 22(6): 1833-1847. https://doi.org/10.1007/s10207-023-00721-y

[8] Zaidi, A.Z., Chong, C.Y., Parthiban, R., Sadiq, A.S. (2022). A framework of dynamic selection method for user classification in touch-based continuous mobile device authentication. Journal of Information Security and Applications, 67: 103217. https://doi.org/10.1016/j.jisa.2022.103217

[9] Bansal, P., Ouda, A. (2024). Continuous authentication in the digital age: An analysis of reinforcement learning and behavioral biometrics. Computers, 13(4): 103. https://doi.org/10.3390/computers13040103

[10] Yang, L., Li, C., You, R., Tu, B., Li, L. (2021). TKCA: A timely keystroke-based continuous user authentication with short keystroke sequence in uncontrolled settings. Cybersecurity, 4(1): 13. https://doi.org/10.1186/s42400-021-00075-9

[11] Shuwandy, M.L., Alasad, Q., Hammood, M.M., et al. (2025). A robust behavioral biometrics framework for smartphone authentication via hybrid machine learning and TOPSIS. Journal of Cybersecurity and Privacy, 5(2): 20. https://doi.org/10.3390/jcp5020020

[12] Baig, A.F., Eskeland, S. (2021). Security, privacy, and usability in continuous authentication: A survey. Sensors, 21(17): 5967. https://doi.org/10.3390/s21175967

[13] Stragapede, G., Vera-Rodriguez, R., Tolosana, R., Morales, A. (2023). BehavePassDB: Public database for mobile behavioral biometrics and benchmark evaluation. Pattern Recognition, 134: 109089. https://doi.org/10.1016/j.patcog.2022.109089

[14] Attrapadung, N., Hanaoka, G., Kotoi-Xie, H.M., et al. (2023). Two-dimensional dynamic fusion for continuous authentication. In 2023 IEEE International Joint Conference on Biometrics (IJCB), Ljubljana, Slovenia, pp. 1-9. https://doi.org/10.1109/IJCB57857.2023.10448866

[15] Li, J., Yi, Q., Lim, M.K., Yi, S., Zhu, P., Huang, X. (2024). MBBFAuth: Multimodal behavioral biometrics fusion for continuous authentication on non-portable devices. IEEE Transactions on Information Forensics and Security, 19: 10000-10015. https://doi.org/10.1109/TIFS.2024.3480363

[16] Zeeshan, N., Bakyt, M., Moradpoor, N., La Spada, L. (2025). Continuous authentication in resource-constrained devices via biometric and environmental fusion. Sensors, 25(18): 5711. https://doi.org/10.3390/s25185711

[17] Stragapede, G., Delgado-Santos, P., Tolosana, R., Vera-Rodriguez, R., Guest, R., Morales, A. (2024). TypeFormer: Transformers for mobile keystroke biometrics. Neural Computing and Applications, 36(29): 18531-18545. https://doi.org/10.1007/s00521-024-10140-2

[18] Momeni, S., BabaAli, B. (2025). Free-text keystroke authentication using transformers: A comparative study of architectures and loss functions. Soft Computing, 29(2): 1259-1272. https://doi.org/10.1007/s00500-025-10524-z

[19] Wang, C., Li, P., Lin, Y. (2025). TSFN: A time-series fusion network for keystroke authentication. The Journal of Supercomputing, 81(14): 1305. https://doi.org/10.1007/s11227-025-07797-4

[20] Wang, C., Li, P., Lin, Y. (2026). Kaformer: Efficient transformer for keystroke authentication. Complex & Intelligent Systems, 12: 249. https://doi.org/10.1007/s40747-026-02385-2

[21] Zhang, Z., Papaioannou, M., Choudhary, G., Dragoni, N. (2026). Client-side continuous authentication using keystroke dynamics: A lightweight pipeline and cross-session evaluation. Electronics, 15(11): 2325. https://doi.org/10.3390/electronics15112325

[22] Sitova, Z., Sedenka, J., Yang, Q., et al. (2016). HMOG: New behavioral biometric features for continuous authentication of smartphone users. IEEE Transactions on Information Forensics and Security, 11(5): 877-892. https://doi.org/10.1109/TIFS.2015.2506542

NOMENCLATURE

$ACER(\tau)$	average classification error rate at threshold τ, dimensionless
$B(\tau)$	deterministic tie-break tuple for threshold selection
CV	coefficient of variation, dimensionless
d_i	movement distance, screen-coordinate units
e_t, e_k	touch/keystroke encoder outputs, dimensionless
$E(\cdot)$	shared dense feature projection
F	final fused vector, dimensionless
$FAR(\tau)$, $FRR(\tau)$	false acceptance/rejection rates, dimensionless
$Flight_i$	press-to-press interval, native time units
$G(q)$	quality/availability gating network
$Hold_j$	key hold duration, native time units
$Intensity_i$	touch intensity, pressure × coordinate/time units
KDS	Key_Diversity_Speed
m_t, m_k	touch/keystroke availability indicators, binary
P	feature-index embedding, dimensionless
$Pressure_i$	touch pressure, native sensor units
q_t, q_k	touch/keystroke quality scores, dimensionless
q	quality/availability vector [q_t, q_k, m_t, m_k], dimensionless
$RhythmRatio$	median-hold/median-flight ratio, dimensionless
s	deterministic neural random seed, integer
$SpatialPressure_i$	spatial pressure, pressure per coordinate unit
$T(\cdot)$	transformer encoder
t_i, Δt_i	event time and interval, native time units
u, f	target-user/fold identifiers, integers
$Unique_KeyID$	number of unique key identifiers, count

v_i	event velocity, coordinate/time units
w_t, w_k	touch/keystroke gate weights, dimensionless
$\tilde{w}$	masked modality-weight vector, dimensionless
X	preprocessed scalar feature token, dimensionless
x_i, y_i	touch-screen coordinates, screen-coordinate units
Z	encoded token representation, dimensionless

Greek symbols

α_j	normalized modality weight, dimensionless
ε	numerical-stability constant, dimensionless
τ, τ^*	candidate/selected decision thresholds, dimensionless

Subscripts

i, j	event and key-event indices
t, k	touch and keystroke modality indices
u, f	target-user and fold identifiers

Superscripts

*	selected/optimized value