Few-Shot Continuous Authentication with Physical Keystrokes

Adam Wherrett — BSc Cyber Security & Digital Forensics

Leeds Beckett University | April 2026

THE PROBLEM

Traditional authentication only verifies identity at sign-in. Sessions remain active for hours or weeks, leaving systems **vulnerable to session hijacking, unauthorised physical access, and credential theft.**

Continuous Authentication (CA) passively verifies users throughout a session — but current Keystroke Dynamics (KD) systems require **hundreds of samples per user** to enrol, making them impractical for real-world use.

RESEARCH QUESTIONS

- Can metric learning achieve competitive accuracy with **<50 enrolment samples**?
- To what extent can the model handle **behavioural drift**?
- How does it compare to deep learning approaches in current research?

WHY KEYSTROKE DYNAMICS?

- Uses existing hardware — standard keyboard only.
- Unique typing rhythms are difficult to imitate.
- Non-intrusive: operates passively in the background.
- Free-text KD enables natural, uninterrupted use.

DATASET & PREPROCESSING

Aalto ~136M Dataset — the largest public keystroke dynamics dataset.

- **168,608** users
- **136M** keystrokes (transcription typing)
- **~3.2M** feature windows after preprocessing

Features: Hold time & flight time

Windowing: Size 20, stride 5 (session z-score normalised)

Split: 80:20 at session level (prevents data leakage)

Storage: HDF5, chunked for efficient streaming

OBJECTIVES & OUTCOMES

Core RQ: Can metric learning achieve competitive EER for free-text keystroke authentication with fewer than 50 enrolment samples per user?

Objective	Status
Design & develop LSTM + triplet loss pipeline on 168K users	✓
Few-shot enrolment evaluation (<50 samples/user)	◉ Partial
Benchmark against current deep learning KD research	✓
Behavioural drift assessment	✗

Why partial? Baseline EER (31.35%) was too high for few-shot evaluation to yield meaningful curves — the comparison would only confirm what the baseline already showed.

Why behavioural drift remains unanswered: Aalto sessions are single-sitting transcription, not longitudinal free-text. No public dataset currently bridges this gap at training scale.

SYSTEM DESIGN

1. **Feature Extraction:** Hold time & flight time per keystroke
2. **Normalisation & Windowing:** Session z-score (size=20, stride=5)
3. **Split:** 80:20 session-level (prevents data leakage)
4. **Storage:** HDF5 (chunked for efficient training)

MODEL ARCHITECTURE

Two-layer **LSTM** network producing 128-dimensional embeddings. Trained with **triplet loss** to learn a metric space where same-user samples cluster together.

Input
(batch, 20, 2)
↓
LSTM Layer 1
128 units, return_sequences=True
↓
LSTM Layer 2
128 units, return_sequences=False
↓
Dropout
rate = 0.5
↓
Dense Projection
128 units, linear activation
↓
Embedding Output
(batch, 128)

TRIPLET LOSS TRAINING

Each training step processes triplets: **Anchor** (user A), **Positive** (same user A), **Negative** (different user B).

Anchor - Positive	Anchor - Negative
Minimise distance	Maximise distance

$$\text{Loss} = \max(0, d(a,p) - d(a,n) + \text{margin})$$

- **Mining:** Random negative sampling (30,000 triplets/epoch)
- **Margin (α):** 1.0
- **Optimiser:** Adam (lr=1e-4, reduce on plateau)

KEY DESIGN DECISIONS

Each choice driven by **few-shot enrolment** (<50 samples) and **open-set recognition** (reject unknown impostors).

Approach	Core Limitation	Verdict
Classification	Fixed output classes → retraining required. Softmax gives class membership, **no distance metric** for thresholding.	✗
Contrastive Loss	Absolute margin requires per-feature tuning; poor choice causes **embedding collapse**. Less sample-efficient.	✗
Proxy-NCA	Proxies **fixed at training time**. Cannot represent unknown impostors — conflicts with open-set requirements.	✗
Triplet Loss	**Selected.** Relative margin → no per-user tuning. New users enrolled via embedding comparison — zero retraining.	✓

Mining Strategy

Strategy	Trade-off	Verdict
Semi-hard	Selects informative triplets within margin. Best sample efficiency but requires **expensive distance calculation** at scale.	Deferred
Random	**Selected.** Low overhead, scales to 3.2M samples. Many triplets contribute zero loss → slower convergence.	✓
Hard	Always picks closest negative → strongest signal but **training instability** risk.	✗

Summary: Triplet loss with random mining was the only combination satisfying **open-set recognition** within available compute — one training run on 3.2M samples was the hard constraint.

EVALUATION METHODOLOGY

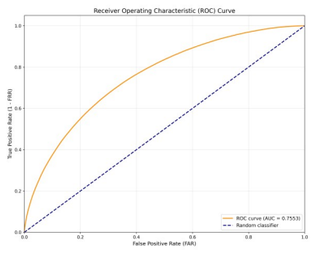
Following Sugrim et al. (2019), no single metric is sufficient. Multiple complementary metrics characterise authentication performance:

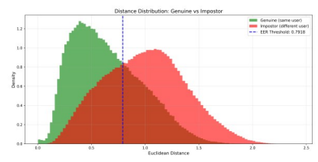
- **ROC AUC** — discriminative ability across all thresholds. AUC > 0.5 confirms learned structure.
- **EER** — threshold where false accepts equal false rejects. A single-point summary statistic.
- **d-prime (d')** — standardised difference between genuine and impostor distance distributions. Measures separability *independent* of any decision threshold.
- **FAR@FRR=1%** — false accept rate at a practical security threshold.

Protocol: Embed test samples → compute pairwise Euclidean distances → genuine vs impostor distributions → ROC sweep (1,000 thresholds) → EER & ROC AUC.

PERFORMANCE RESULTS

0.755	31.35%
ROC AUC	EER
0.986	**89.05%**
d-prime	FAR@FRR=1%





ROOT CAUSES & LIMITATIONS

While the model works, the **31.35% EER** is substantially higher than the 2.67–10% range in current literature.

Limited Features: Only hold time & flight time. N-gram features (digraph/trigraph latency) are more discriminatory.

Random Mining: Many training triplets contributed zero loss. Semi-hard mining would improve sample efficiency, but was computationally expensive at scale.

Dataset Constraints: Aalto captures transcription typing (not free composition), and sessions are single-sitting — limiting behavioural drift assessment.

FUTURE WORK

Discriminative Features: Incorporate digraph and trigraph latency features alongside hold and flight times to capture higher-order typing patterns.

Semi-Hard Mining: Replace random negative sampling with semi-hard triplet mining to increase the proportion of informative training triplets.

Longitudinal Evaluation: Evaluate on datasets with multi-session data collected over time to properly assess drift resistance in a few-shot setting.

CONNECT

Scan for full thesis & portfolio