Journal of Behavioral Data Science, 2027, *7* (1), 1-29.
DOI: https://doi.org/10.35566/jbds0228

Typing Rhythms Beyond the Latin Script: An Amharic–English Keystroke Dynamics Benchmark

Animaw Kerie Aseres[*1,3], Lemlem Kassa Tegegne[1,3], Wongel Dawit Daniel[4], and Asrat Mulatu Beyene[2,3]

[1] Department of Software Engineering, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia
[2] Department of Electrical and Computer Engineering, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia
[3] High Performance Computing and Big Data Analytics Center of Excellence, AASTU, Addis Ababa, Ethiopia
[4] Languages and Spatial Technologies Lab, Carnegie Mellon University Africa, Kigali, Rwanda
[*] Corresponding author: Animaw Kerie Aseres, animaw.kerie@kue.edu.et

Abstract. Keystroke dynamics—the timing of a person's key presses and releases—is a low-cost, non-intrusive behavioral biometric. However, public benchmarks and most empirical evidence focus on typing in Latin scripts, predominantly English, and do not cover speakers of Ge'ez-script (Ethiopic) languages such as Amharic, which has more than 35 million native speakers. Ethiopic syllabic characters are commonly entered through multi-key transliteration, creating different motor-planning demands. This raises the question of whether timing-based verification methods developed for Latin scripts transfer effectively to other writing systems. We present and characterize a keystroke dynamics corpus collected from a shared pool of typists completing three text-entry tasks—fixed text, free text, and password entry—in Amharic and English. The corpus contains 114,955 labeled keystroke events across five sub-corpora (4,751–51,813 rows each), with up to 25 raw and derived timing fields per event. We report a systematic exploratory data analysis (EDA) covering structural composition; missingness and duplication; demographic and skill-level coverage; univariate and bivariate timing distributions; feature redundancy (correlation and hierarchical clustering); per-user sample imbalance; and user distinguishability in Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) projections. The analysis finds strongly right-skewed timing features, a small number of algebraically redundant feature pairs, a coefficient of variation in per-user sample counts of up to 2.4, and substantial visual overlap between users in low-dimensional projections. Together, these properties plausibly explain the 40–60% rank-1 accuracy and 10–20% equal error rate (EER) previously obtained with baseline classifiers on this cohort. Based on these diagnostics and recent literature, we recommend feature engineering, session- and user-aware evaluation protocols, and careful model selection for keystroke biometrics in low-resource,

non-Latin-script settings. We also make the EDA pipeline available for benchmarking. To the best of our knowledge, this is the first analysis of tri-modal keystroke dynamics data in an Ethiopic-script language, providing a reference point for cross-lingual and cross-script research in behavioral biometrics.

Keywords: Keystroke dynamics · Behavioral biometrics · Amharic keystroke · · Continuous authentication · Ethiopic script

1 Introduction

Behavioral biometrics authenticates or continuously verifies a user's identity based on how an action is performed, rather than on static physiological traits (Abuhamad, Abusnaina, Nyang,x& Mohaisen, 2021; Stylios, Kokolakis, Thanou,x& Chatzis, 2021). Keystroke dynamics, which measures how long keys are held and the intervals between successive key events, is particularly attractive: it requires no special hardware, can be captured passively during ordinary typing, and supports both one-time login verification and continuous verification after login (Shadmanxet al., 2025). The field has advanced from distance-based and one-class anomaly detectors for short, fixed passwords (Killourhyx& Maxion, 2009) to instance-based statistical matchers for free text (Ayotte, Banavar, Hou,x& Schuckers, 2020) and deep sequence models, including recurrent networks trained with metric-learning losses (Acien, Morales, Monaco, Vera-Rodriguez,x& Fierrez, 2022), Transformers (Stragapede, Delgado-Santos,xet al., 2023, 2024), and hybrid recurrent–convolutional architectures with attention (González, Stragapede, Vera-Rodriguez,x& Tolosana, 2025). These approaches have been evaluated on populations of up to hundreds of thousands of users.

Nevertheless, empirical evidence remains concentrated on English and other Latin-script languages typed on QWERTY layouts (Shadmanxet al., 2025). Widely used benchmarks—including the CMU fixed-password dataset (Killourhyx& Maxion, 2009), the GREYC database (Giot, El-Abed,x& Rosenberger, 2009), the Buffalo dataset (Sun, Ceker,x& Upadhyaya, 2016), the Clarkson II corpus (described in Ayottexet al., 2020), and Aalto University typing data (Dhakal, Feit, Kristensson,x& Oulasvirta, 2018)—which underpin recent large-scale benchmarks (Stragapede, Vera-Rodriguez,xet al., 2024; Stragapedexet al., 2025)—contain no non-Latin scripts. The few available multilingual studies suggest this is more than a cosmetic limitation: verification performance declines when enrollment and test languages differ (Wahabxet al., 2022), and typing proficiency and consistency vary across a user's languages and alphabets (Kang, 2015; Kimx& Kang, 2020). Ge'ez-script (Ethiopic) languages are a pertinent but understudied case. Amharic, the Ethiopian federal government's main working language, has more than 35 million native speakers and more than 25 million second-language speakers (Indiana University, n.d.), yet remains a low-resource language for language technology (Tonjaxet al., 2024; Yimam, Ayele, Venkatesh, Gashaw,x& Biemann, 2021). Its script is an abugida: each character (*fidel*) represents a consonant–vowel syllable. It has 34 base consonants, each realized in seven vowel forms, plus labialized variants (Belayxet al., 2020). On standard keyboards, users typically enter these characters through phonetic transliteration methods, where a syllable may require one or more Latin keystrokes (Microsoft, n.d.). These input methods create motor-planning, key-sequencing, and visual-search demands that differ from direct alphabetic typing. Only a purpose-built corpus can determine whether keystroke-timing biometrics generalize to this setting and whether models built on Latin-script assumptions apply.

A second gap concerns modality coverage within a single cohort. Researchers have typically examined fixed-text, free-text, and password-style typing separately, with different populations, protocols, and feature-extraction schemes, making cross-modality comparisons difficult (Shadmanxet al., 2025). Free-text data, in particular, exhibit greater intra-subject variability than fixed-text data because content, language, and the user's state influence typing behavior (Acienxet al., 2022; Stragapede, Delgado-Santos,xet al., 2023). Wahabxet al. (2022) released multi-keyboard, bilingual (English–Chinese) data to support more controlled comparisons, but such resources remain rare. Here, we provide a corpus in which participants from a shared pool contribute data across all three modalities, allowing us to examine how text constraints and content shape timing-feature distributions and user separability.

We address these gaps by introducing a multimodal keystroke dynamics corpus comprising five sub-corpora: Amharic fixed-text, English fixed-text, Amharic free-text, English free-text, and password text (Latin letters, digits, and symbols). We collected data from a shared pool of typists who self-reported their gender and typing skill. Rather than proposing a new classifier, we conduct a comprehensive exploratory data analysis (EDA) to examine the corpus's structure, quality, and statistical properties before training any models. This step is often underreported in keystroke biometrics, even though data properties—including dataset size, per-user sample counts, and dependence between training and test samples—strongly influence reported performance (Kapoorx& Narayanan, 2023; Wahabx& Hou, 2023). In preliminary baseline experiments with the same cohort, standard classifiers achieved only 40–60% rank-1 accuracy and 10–20% Equal Error Rate (EER). These results are considerably worse than those reported for large, English-dominant free-text benchmarks, such as TypeNet's 2.2% EER on physical keyboards (Acienxet al., 2022) and the approximately 3.3% EER achieved by the best desktop systems in the KVC-onGoing challenge (Stragapedexet al., 2025). Because protocols, data volumes, and populations differ, these figures are not directly comparable. Nevertheless, the gap motivates a careful diagnosis of the corpus itself.

1.1 Contributions

The contributions of this work are as follows:

- To the best of our knowledge, we present the first keystroke dynamics corpus to include an Ethiopic-script language (Amharic) alongside English data from the same participant pool, spanning fixed-text, free-text, and password-text paradigms.
- We document a complete EDA workflow—structural, statistical, and manifold-based—and report its outputs as baseline diagnostics to support future research using the corpus.
- We connect the EDA findings quantitatively and qualitatively to previously obtained baseline accuracy and EER, showing how feature skewness, redundancy, per-user imbalance, and overlapping user manifolds may each contribute to the observed performance ceiling.
- We offer practical recommendations for feature engineering, evaluation protocol design, and model selection in low-resource, non-Latin-script keystroke biometrics. These recommendations draw on recent work in deep and statistical keystroke authentication, mobile and continuous authentication, and privacy and fairness in behavioral biometrics.

The remainder of the paper is organized as follows. Section section2 reviews related work on keystroke dynamics models, evaluation protocols, mobile and multimodal biometrics, the effects of language and script, benchmark datasets, and robustness, privacy, and fairness. Section section3 describes the data collection, feature schema, and EDA methodology, and Section section4

presents the EDA results. Section section5 discusses their implications for modeling and script-inclusive behavioral biometrics. Section section6 outlines limitations and threats to validity, and Section section7 concludes.

2 Related Work

2.1 Deep and statistical models for keystroke dynamics

Recent research has largely moved from hand-crafted distance metrics to learned representations. Acienxet al. (2022) trained LSTM networks (TypeNet) with softmax, contrastive, and triplet losses on the Aalto databases (about 168,000 desktop and 60,000 mobile users). They reported EERs of 2.2% and 9.2% for physical and touchscreen keyboards, respectively, with only a moderate increase in error when the test population grew to 100,000 subjects. Transformer-based models followed. Stragapede, Delgado-Santos,xet al. (2023) introduced a Transformer for free-text mobile keystrokes, achieving a 3.84% EER with five enrollment sessions; TypeFormer lowered this to 3.25% using five enrollment sessions of 50 keystrokes each (Stragapede, Delgado-Santos,xet al., 2024). Type2Branch combines parallel recurrent and convolutional branches with attention mechanisms and a set-based loss, achieving the lowest errors in the KVC-onGoing challenge (Gonzálezxet al., 2025). Metric-learning designs, which bring samples from the same user closer together and push samples from different users apart, are also attractive when only a few samples per user are available. For example, Lis, Niewiadomska-Szynkiewicz,xand Dziewulska (2023) used a Siamese network with n-shot classification to verify users typing partial passwords. However, these advances have been demonstrated mainly on large Latin-script corpora. On smaller datasets, simple statistical methods can outperform deep models: Wahabxand Hou (2023) reported EERs of 1.3% for instance-based tail area density (ITAD) and 4.3% for scaled Manhattan distance, compared with 19.18% for TypeNet, on a small dataset; TypeNet, however, performed better on two larger datasets. This finding is directly relevant to the present corpus, whose sub-corpora contain 34–183 users and have heavily tailed per-user sample counts (Section subsection4.7).

2.2 Free-text, continuous authentication, and evaluation protocols

Free-text keystroke dynamics is more challenging than fixed-text analysis because content, vocabulary, and the user's emotional and physical state influence typing rhythm, increasing intra-subject variability (Acienxet al., 2022; Stragapede, Delgado-Santos,xet al., 2023). Instance-based statistical approaches address this challenge by comparing individual digraph timings with a user profile. With 200 digraphs, ITAD achieved EERs of 7.8% on the uncontrolled Clarkson II corpus and 3.0% on the Buffalo dataset (Ayottexet al., 2020). Reported results also depend heavily on evaluation design. The Keystroke Verification Challenge (KVC) introduced a standardized protocol using tweet-length sequences from more than 185,000 subjects (Stragapede, Vera-Rodriguez,xet al., 2024). Its latest version reported best EERs of about 3.33% for desktop and 3.61% for mobile (Stragapedexet al., 2025). The winning system's mean per-subject EERs (0.77% desktop and 1.03% mobile) rose to about 3.25% and 3.61%, respectively, when a single global decision threshold was applied (Gonzálezxet al., 2025). This illustrates how strongly reported error rates depend on threshold and averaging choices. Our analysis of per-user imbalance (Section subsection4.7) relates directly to these issues: it quantifies the heterogeneity of user contributions in the present corpus before any adaptation or calibration.

2.3 Mobile, touch, and multimodal behavioral biometrics

On smartphones, soft-keyboard timing is noisy, so researchers often combine keystroke features with touch, motion, and orientation signals, as in the HMOG feature set (Sitováxet al., 2016). Kimxand Kang (2020) combined keystroke and sensor features collected as participants typed free text in English and Korean, reporting EERs below 1% with a single reference set. Public mobile databases such as BehavePassDB, which records touchscreen and background-sensor data across several everyday tasks, now support standardized benchmarking (Stragapede, Vera-Rodriguez, Tolosana,x& Morales, 2023). The associated MobileB2C competition showed that passive mobile authentication is feasible but remains challenging (Stragapedexet al., 2022). Surveys of sensor-based and behavioral continuous authentication describe the range of fused modalities—including keystroke, touch, gait, motion, and voice—and persistent challenges in usability, performance, and adoption (Abuhamadxet al., 2021; Styliosxet al., 2021). The project from which the present data are derived ("Multimodal Biometric Fusion") shares this long-term goal: its fixed-text, free-text, and password sub-corpora are designed to support the addition of other modalities in future work.

2.4 Language, script, and input-method effects

Evidence on the effects of language and script on keystroke biometrics is limited but consistent. Kang (2015) studied 83 Korean participants typing in Korean and English. They were more proficient but less consistent in Korean, their primary language and the alphabet they knew better. Which factor mattered most depended on whether authentication used only latencies or also key-sequence information. Wahabxet al. (2022) collected bilingual data from 26 participants fluent in English and Chinese and found substantial cross-language degradation: enrolling in Chinese and testing in English raised the EER from 2% to 16%, while the reverse raised it from 3.3% to 9.7%. These findings suggest that keystroke templates are partly language-specific and that differences in users' proficiency across languages shape timing distributions. The Ethiopic script presents an additional challenge. Amharic characters are syllabic: each of 34 base consonants has seven vowel forms, plus labialized forms (Belayxet al., 2020). Phonetic input methods map each syllable to one or more Latin keystrokes, so the number of keystrokes per character varies (Microsoft, n.d.). Amharic is also under-resourced in language technology more broadly, although semantic models and multilingual language models for Ethiopian languages are beginning to emerge (Tonjaxet al., 2024; Yimamxet al., 2021). To our knowledge, no previous keystroke dynamics study has examined typing in the Ethiopic script.

2.5 Dataset benchmarks and reproducibility

Progress in the field has been supported by public corpora, including the CMU fixed-password database (Killourhyx& Maxion, 2009), the GREYC keystroke database (Giotxet al., 2009), the Buffalo dataset (Sunxet al., 2016), the Clarkson II corpus of uncontrolled free text (Ayottexet al., 2020), and Aalto University's large-scale typing data (Dhakalxet al., 2018). With data from more than 168,000 desktop participants, the Aalto collection underpins current deep-learning benchmarks (Acienxet al., 2022; Stragapede, Vera-Rodriguez,xet al., 2024). More recent releases include mobile multimodal data (Stragapede, Vera-Rodriguez,xet al., 2023), multi-keyboard and bilingual English–Chinese data (Wahabxet al., 2022), and the KeyRecs dataset, which includes fixed- and free-text samples from 100 participants representing 20 nationalities (Dias, Vitorino,

Maia, Sousa,x& Praça, 2023). A recent survey of keystroke datasets and algorithms identifies only a few multilingual collections, including the English–Korean FACT data (Kimx& Kang, 2020) and English–Chinese bilingual data (Wahabxet al., 2022), and finds no collections covering an Ethiopic script (Shadmanxet al., 2025). Table table1 places the present corpus alongside representative benchmarks.

Table 1. Positioning of the present multimodal Amharic–English corpus relative to representative keystroke-dynamics benchmark datasets reported in the literature

Dataset	Script / Language	Modality	Approx. users	Reported error regime
CMU Fixed-Password (Killourhyx& Maxion, 2009)	Latin / English	Fixed-text (password)	51	∼10% EER (classical distance metrics)
GREYC Keystroke (Giotxet al., 2009)	Latin / English, French	Fixed-text (password)	133	∼5–10% EER
Buffalo Free-Text (Sunxet al., 2016)	Latin / English	Free-text	148	Session/day effects reported; EER varies widely
Clarkson II (see Ayottexet al., 2020)	Latin / English	Free-text (longitudinal)	103	Long-term drift documented
Aalto Typing (Dhakalxet al., 2018)	Latin / English (web-scale)	Free-text	∼168,000	Large-scale rhythm statistics only
HMOG (Sitováxet al., 2016)	Latin / English (mobile)	Fixed + free (touch fused)	100	Multimodal fusion reduces EER vs. keystroke alone
Present corpus	Ethiopic (Amharic) + Latin (English)	Fixed-text, free-text, password (same cohort)	34–183 per sub-corpus	40–60% accuracy / 10–20% EER (baseline, pre-optimization)

Table table1 highlights two distinctive features of the present study: (i) to our knowledge, it is the first to include a language written in Ethiopic script, and (ii) it is among the few corpora to collect fixed-text, free-text, and password typing from the same participant pool, enabling cross-modality comparisons that are difficult with separately collected datasets. The error regimes listed

in Table table1 were measured under heterogeneous protocols and are provided for context only, not for like-for-like comparison.

2.6 Robustness, privacy, and fairness in behavioral biometrics

Beyond accuracy, recent work has also examined robustness, privacy, and fairness. Carefully designed adversarial attacks can target behavioral biometric models (Styliosxet al., 2021), and synthetic keystroke timings generated from population- or user-specific statistics can substantially increase false acceptance rates in existing verifiers, underscoring the need for dedicated liveness detection (González, Calot, Ierache,x& Hasperué, 2022). Typing data also reveal information beyond identity: keystroke features can classify users by age and gender (Tsimperidis, Yucel,x& Katos, 2021), and mobile sensor and keystroke data can expose demographic, health, and behavioral attributes (Delgado-Santosxet al., 2022). In the KVC benchmark, replacing key-content features with timing-only features preserved satisfactory performance while reducing privacy intrusion (Stragapede, Vera-Rodriguez,xet al., 2024). Demographic bias is a recognized challenge across biometric modalities (Drozdowski, Rathgeb, Dantcheva, Damer,x& Busch, 2020). For keystroke dynamics, KVC introduced the Skewed Impostor Ratio to quantify group-level bias (Stragapede, Vera-Rodriguez,xet al., 2024), and the ongoing challenge found that verification scores reflect subjects' age and gender to varying degrees (Stragapedexet al., 2025). Our EDA reports gender and skill-level coverage for every sub-corpus (Section subsection4.3), enabling future fairness analyses of this corpus to account for subgroup composition.

3 Materials and Methods

3.1 Participant cohort and data-collection protocol

We collected keystroke data from volunteer typists using a custom Android application that recorded key-press and key-release times to the nearest millisecond. During enrollment, participants reported their gender and typing proficiency (e.g., Novice, Intermediate, Advanced, or Expert). We labeled each submission with a session number and, for fixed-text tasks, with its prompt sentence. This design yielded five sub-corpora: fixed- and free-text entry modes in each of two languages (Amharic and English), plus a password-text sub-corpus typed with Latin letters, digits, and special symbols. Because user identifiers remain consistent across the data collection platform, the five sub-corpora are expected to contain overlapping but nonidentical participant groups, with 360–366 user IDs per sub-corpus. This enables the comparisons in Section subsection4.9.

3.2 Feature schema

Each sub-corpus row corresponds to a logged key event or an event-derived record and contains up to 25 fields: participant ID, session ID, self-assessed skillLevel and gender; the target Sentence and completeUserInput (in fixed-text sub-corpora only); a wall-clock Timestamp; the struck key's text and type (keyText, keyType); per-event timings (flightTime, holdTime, interKeyTime, pressTime, and releaseTime); and engineered aggregates calculated at the session level and repeated on every row for that session: AverageKeyPressDuration, AverageKeyReleaseDuration, KeyPressVariability, KeyReleaseVariability, AverageInterkeyTime, KeyPressRate, KeyReleaseRate, AverageSeekTime, TypingSpeed, CharactersPerMinute, HoldFlightRatio, and FlightHoldRatio. Of these fields,

18 in fixed-text sub-corpora and 19 in free-text and password sub-corpora are numeric and were used in the quantitative EDA reported in Section section4. The latter sub-corpora lack a separate Sentence ID field and therefore include one additional numerically encoded field. Hold times and inter-key latencies are standard timing primitives in keystroke dynamics research (Shadmanxet al., 2025).

3.3 Exploratory data analysis pipeline

All analyses were conducted in Python 3. pandas and NumPy were used for data manipulation; seaborn and Matplotlib for statistical graphics; scikit-learn for standardization (StandardScaler), principal component analysis (PCA), and t-distributed stochastic neighbor embedding (t-SNE; van der Maatenx& Hinton, 2008; default perplexity and a fixed random seed of 42); and umap-learn for uniform manifold approximation and projection (UMAP; McInnes, Healy, Saul,x& Großberger, 2018; default parameters and a random seed of 42). The same procedure was applied independently to each of the five sub-corpora and covered: (1) structural analysis (shape, data types, and cardinality of users, sessions, and sentences); (2) a missing-value audit visualized using binary heatmaps; (3) identification of duplicate rows; (4) descriptive statistics for numeric variables (count, mean, standard deviation, and quartiles); (5) demographic and skill-level coverage; (6) univariate distributions of typing speed and three core timing primitives (hold time, flight time, and inter-key time); (7) boxplot-based screening of all numeric variables for outliers in each sub-corpus; (8) Pearson correlation matrices and cluster maps based on agglomerative clustering; (9) feature-level variation and skewness; (10) per-user aggregation of each numeric feature (mean and standard deviation) by participant id; (11) per-user sample and session distributions as indicators of data imbalance; (12) within-user stability (per-user standard deviation of hold time) and session-level variation (per-user mean hold time across sessions); (13) two- and three-dimensional projections of the standardized numeric feature space (PCA, t-SNE, and UMAP), colored by participant identity as an unsupervised proxy for user separability; and (14) keystroke-content analysis (key-frequency and key-type distributions) and comparison of typing speed between free and fixed text. Because t-SNE and UMAP emphasize local neighborhoods and do not reliably preserve global structure, and because two-dimensional embeddings of high-dimensional data inevitably introduce distortion, we interpret the projections in step (13) qualitatively rather than as quantitative measures of separability (Charix& Pachter, 2023; Wang, Huang, Rudin,x& Shaposhnik, 2021).

3.4 Preprocessing decisions

Because several downstream analyses (PCA, t-SNE, UMAP, and correlation) are sensitive to missing values, missing numeric values were replaced with zero or the column mean. Identifier columns were excluded from projection and correlation analyses because they would trivially encode user identity in the feature space—a form of illegitimate-feature leakage (Kapoorx& Narayanan, 2023). Zero-variance rows and columns were also removed before correlation-based clustering and dimensionality reduction. All figures in Section section4 follow this workflow and are presented as composite, multi-panel exhibits, with one panel per sub-corpus. Because imputation choices can affect correlation and projection results, we revisit this issue in Section section6.

4 Results

4.1 Corpus composition

The full dataset contains 114,955 keystroke-event rows across five sub-corpora (Table table2). English fixed-text is the largest, with 51,813 rows, 183 users, and 1,787 sessions. Amharic free-text is the smallest, with 4,751 rows, 34 users, and 239 sessions; this may be because composing unconstrained text in Amharic requires more effort. The fixed-text sub-corpora include few distinct prompt sentences (7 for Amharic and 13 for English), while the free-text and password sub-corpora have no fixed prompts.

Table 2. Structural composition of the five sub-corpora

Sub-corpus	Language	Modality	Rows	Users	Sessions	Unique prompts	Numeric feat.
amharic_fixed	Amharic	Fixed-text	18,684	95	918	7	18
english_fixed	English	Fixed-text	51,813	183	1,787	13	18
amharic_free	Amharic	Free-text	4,751	34	239	n/a	19
english_free	English	Free-text	20,417	75	612	n/a	19
password	English (chars/ digits/ symbols)	Password-text	19,290	121	1,165	n/a	19
Total	—	—	114,955	34–183[a]	4,721	—	—

[a]*Per sub-corpus. User IDs show significant overlap between sub-corpora (maximum observed identifier of ∼360–366 in fixed-text sub-corpora), suggesting a shared pool of participants as opposed to separate pools; it is incorrect to add up user numbers within each sub-corpus to find unique users.*

4.2 Data quality: missingness and duplication

Missing values are uncommon in the fixed-text sub-corpora (0.43–0.59% of all cells) but rise to 4.4–4.7% in the free-text and password sub-corpora (Table table3). This increase reflects two factors: flightTime and interKeyTime cannot be calculated for the first key in each sequence because it has no preceding key, and the free-text tasks lack an explicit Sentence field. Figure figure1 shows binary heatmaps of missing-value locations in each sub-corpus. The structured (banded) pattern in several panels indicates that missingness is not randomly distributed across the data (i.e., it is not missing completely at random), with implications for imputation (Sections subsection5.2 and section6). Duplicate rows are very rare: 6–159 rows, or less than 0.4% of any sub-corpus.

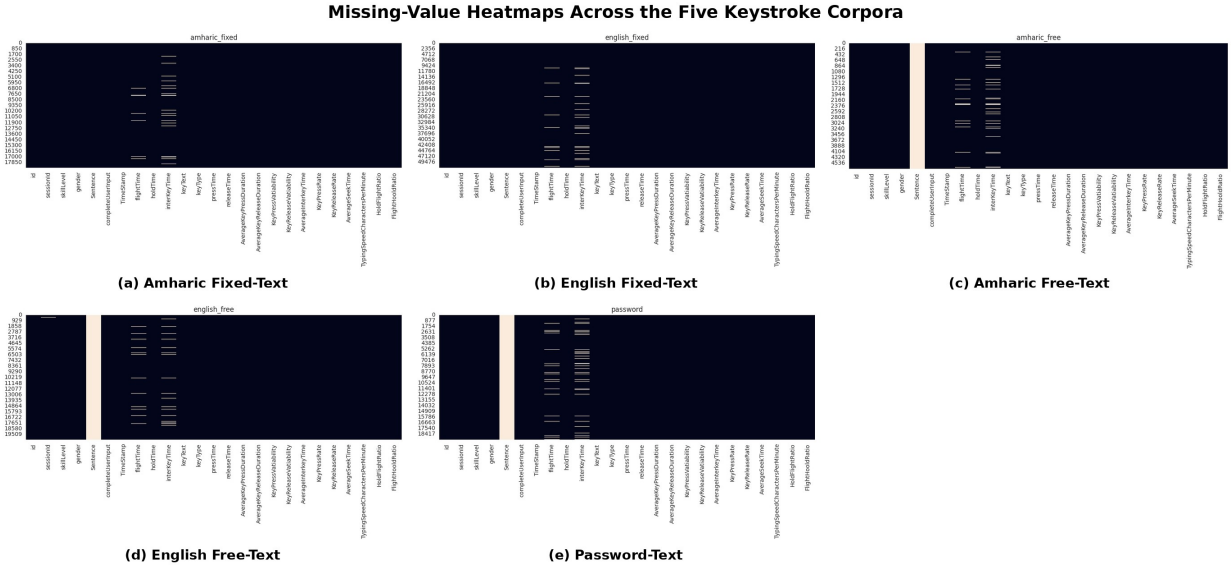


Figure 1. Binary missing-value heatmaps for the five sub-corpora (yellow = missing). Missingness is structured rather than random, concentrated in flight-time, inter-key-time, and sentence-identifier columns.

Table 3. Final data-quality summary (all rows/columns, before any imputation)

Sub-corpus	Rows	Users	Sessions	Missing cells	Missing (%)	Duplicate rows
amharic_fixed	18,684	95	918	2,755	0.59	59
english_fixed	51,813	183	1,787	5,554	0.43	135
amharic_free	4,751	34	239	5,469	4.60	38
english_free	20,417	75	612	22,279	4.36	159
password	19,290	121	1,165	22,786	4.72	6

4.3 Demographic and skill-level coverage

Figure figure2 shows that a single self-reported gender predominates in each sub-corpus. Figure figure3 shows that self-reported skill levels cluster toward the upper end of the scale, with relatively few novices. Both imbalances are relevant to subgroup fairness evaluation and to assessing whether classifiers perform equally well across subgroups (Section subsection2.6; Drozdowskiet al., 2020). Exact subgroup counts are available in the outputs of our EDA pipeline.

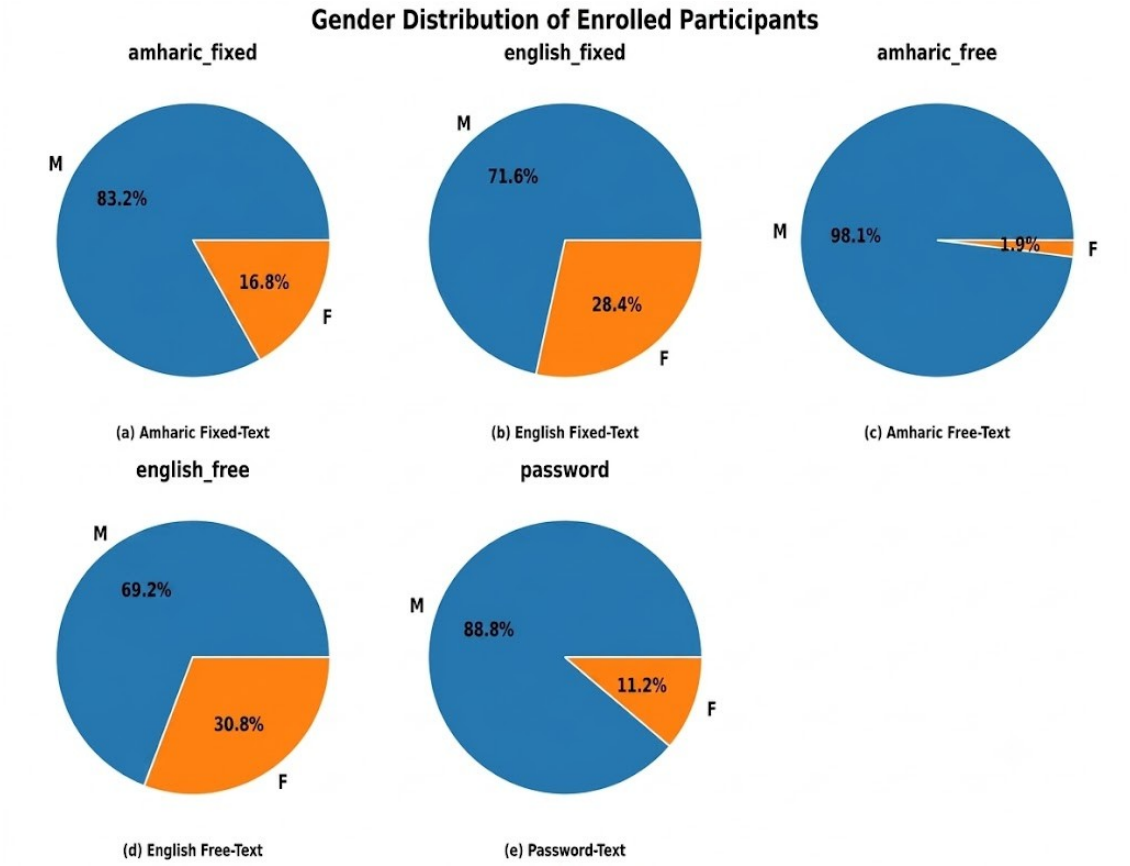


Figure 2. Self-reported gender distribution by sub-corpus. Each sub-corpus is imbalanced, so fairness-oriented evaluation should account for this.

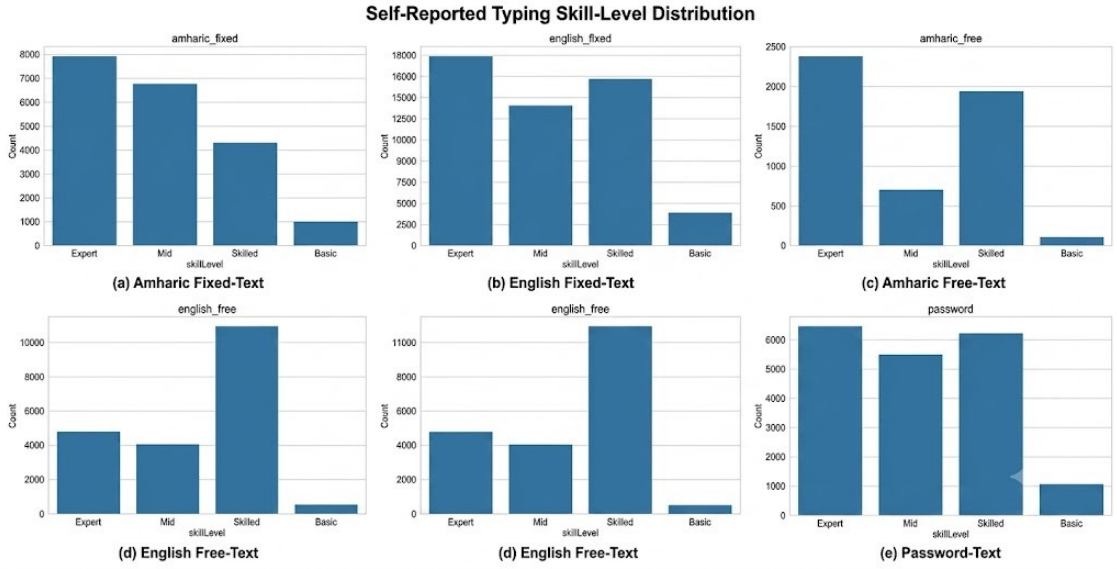


Figure 3. Self-reported typing skill-level distribution by sub-corpus, showing a concentration in higher-skill categories.

4.4 Timing-feature distributions

Table table4 presents the mean and standard deviation for three key timing features—hold time, flight time, and inter-key time, measured in milliseconds—and typing speed, measured in characters per minute (CPM), across each sub-corpus. Two patterns stand out. First, typing speed varies substantially by modality and language. The password sub-corpus has the slowest average speed (33.9 CPM), likely because finding and entering digits and symbols takes extra effort, while English free-text is the fastest (98.6 CPM), in line with natural, fluent composition. Second, the timing features show very large standard deviations relative to their means: flight-time standard deviation exceeds the mean in all five sub-corpora, while inter-key-time standard deviation exceeds the mean in four of the five. The most extreme case is English fixed-text, where flight-time standard deviation is 9,962.6 ms, compared with a mean of 805.7 ms. Figures figure4 and figure5 show the corresponding distributions.

Table 4. Descriptive statistics (mean ± standard deviation) of core timing features, computed over all rows in each sub-corpus

Sub-corpus	Hold time (ms)	Flight time (ms)	Inter-key time (ms)	Typing speed (CPM)
amharic_fixed	93.1 ± 51.6	1,413.7 ± 1,988.6	2,889.1 ± 2,907.8	43.7 ± 25.3
english_fixed	95.7 ± 56.6	805.7 ± 9,962.6	1,695.3 ± 14,363.3	68.0 ± 41.3
amharic_free	80.5 ± 38.3	924.8 ± 1,001.3	1,910.7 ± 1,564.4	71.0 ± 69.3
english_free	89.8 ± 46.8	478.6 ± 1,368.9	1,030.3 ± 2,004.6	98.6 ± 71.3
password	92.7 ± 68.5	1,108.4 ± 2,474.8	2,307.2 ± 3,491.8	33.9 ± 15.4

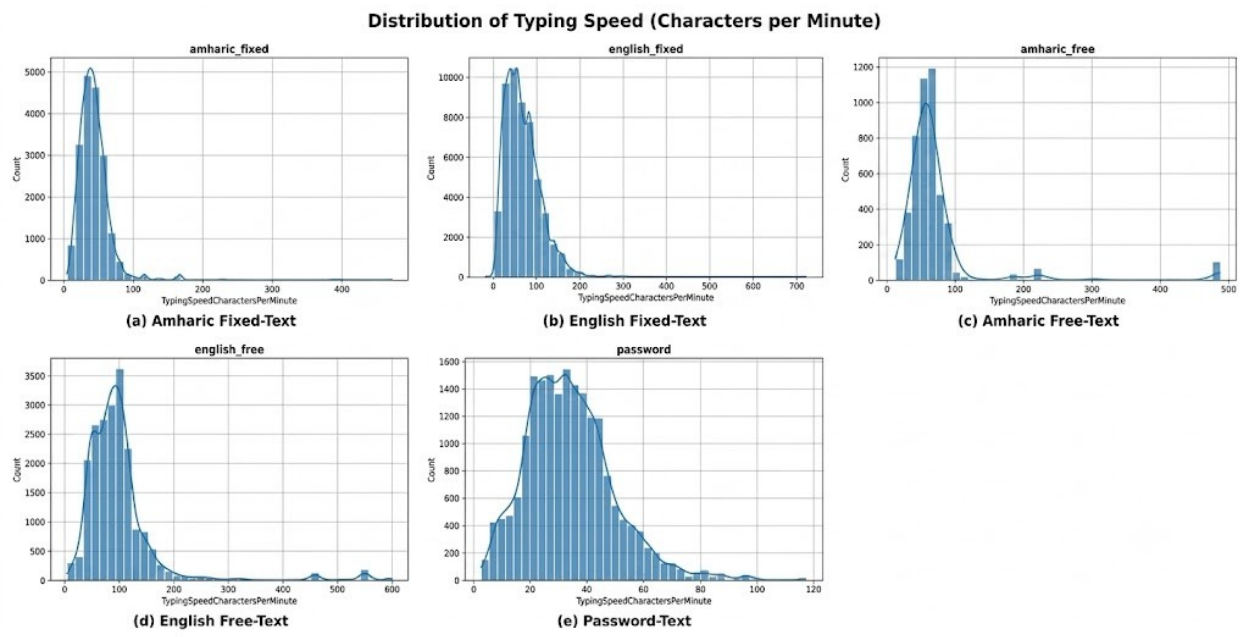


Figure 4. Distribution of typing speed (characters per minute) by sub-corpus, showing a right skew common to all five modalities.

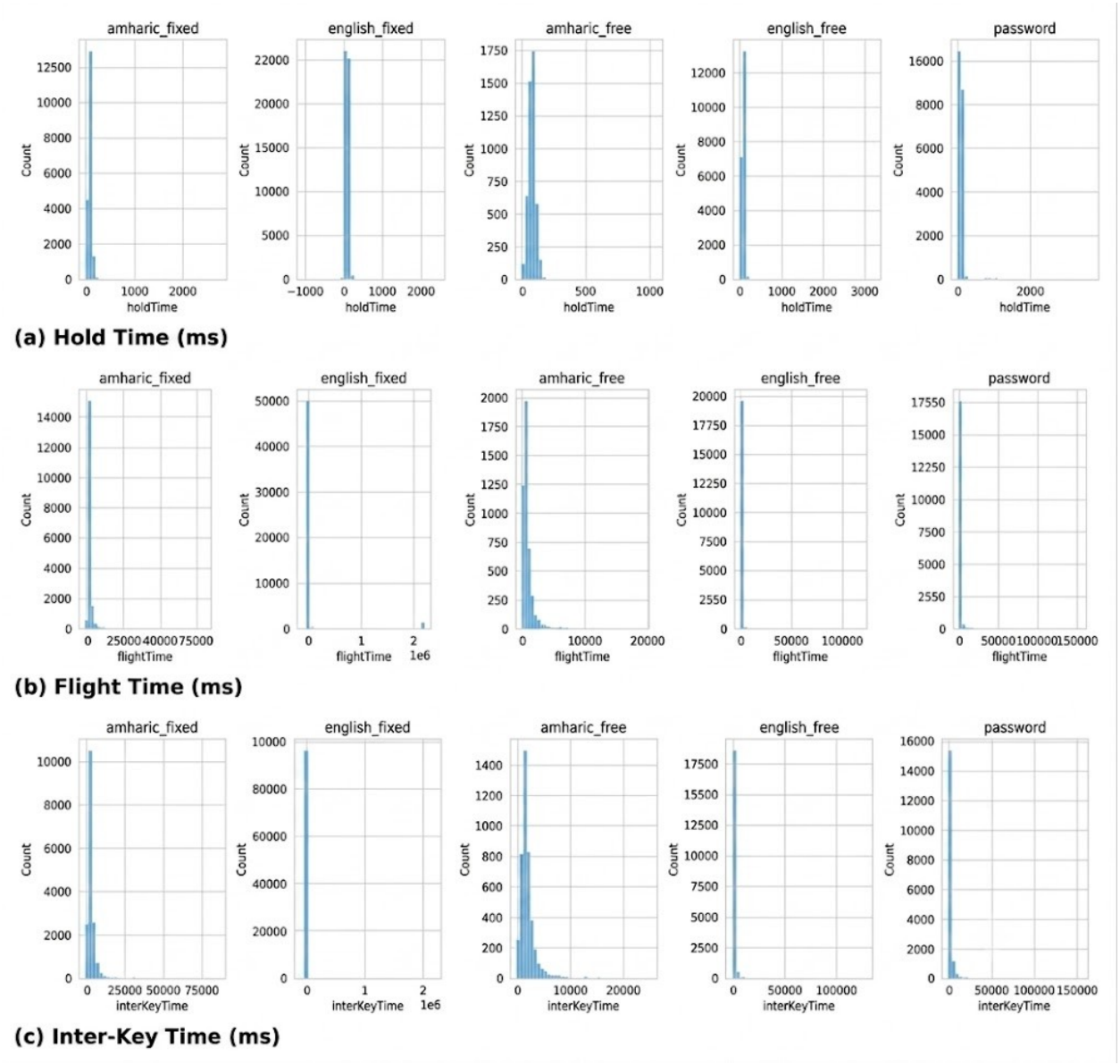
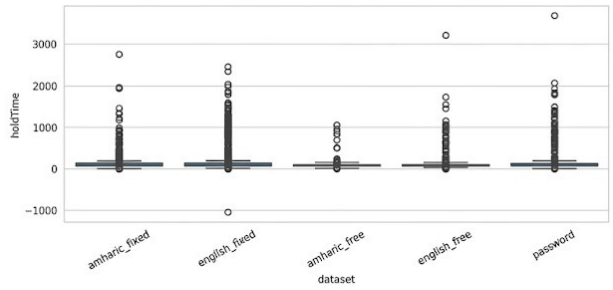


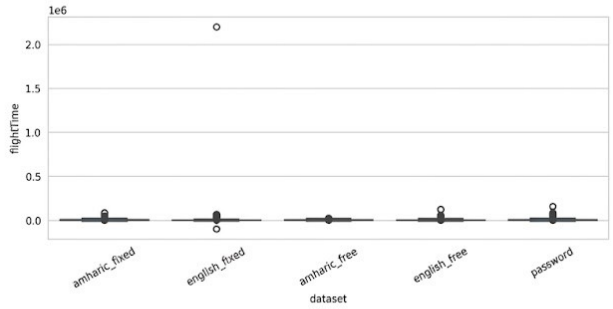
Figure 5. Distributions of (a) hold time, (b) flight time, and (c) inter-key time across sub-corpora, each panel overlaying all five sub-corpora for direct comparison. All three primitives are right-skewed with long upper tails.

4.5 Outlier screening

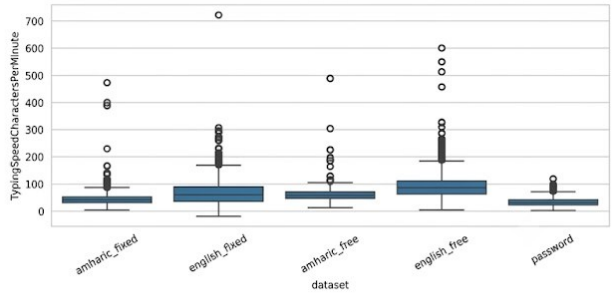
Cross-sub-corpus boxplots (Figure figure6) of hold time, flight time, and typing speed confirm the heavy tails indicated in Table table4. Each feature has a narrow interquartile range and numerous outliers several interquartile ranges above the upper quartile. Some flight-time and inter-key-time values are also negative; we attribute these to clock-skew artifacts rather than genuine negative intervals (see Section section6 for an alternative explanation). We retained these points in the summary statistics instead of silently discarding them for two reasons: (a) long pauses may reflect genuine cognitive or motor events of interest, and (b) any trimming decision should be made transparently during modeling, not embedded in a shared EDA result. We flag the negative values for resolution during data cleaning.



(a) Hold Time (ms)



(b) Flight Time (ms)



(c) Typing Speed (CPM)

Figure 6. Cross-sub-corpus boxplots of (a) hold time, (b) flight time, and (c) typing speed, showing compact interquartile ranges with extensive high-value outliers in every sub-corpus.

4.6 Feature redundancy and correlation structure

The Pearson correlation heatmaps (Figure figure7) and hierarchically clustered correlation maps (Figure figure8) indicate that several engineered features are algebraically or nearly algebraically redundant with the raw features used to derive them. Across all sub-corpora, holdTime correlates perfectly or almost perfectly with AverageKeyPressDuration, while flightTime is strongly correlated with AverageInterkeyTime. AverageSeekTime is also strongly correlated with AverageKeyReleaseDuration, as these session-level aggregates are recalculated for each row in a session. KeyPressRate and KeyReleaseRate are likewise redundant across all five sub-corpora. As a result, the nominal 18–19 numeric columns represent far fewer independent degrees of freedom; the cluster maps suggest approximately 8–10 per sub-corpus.

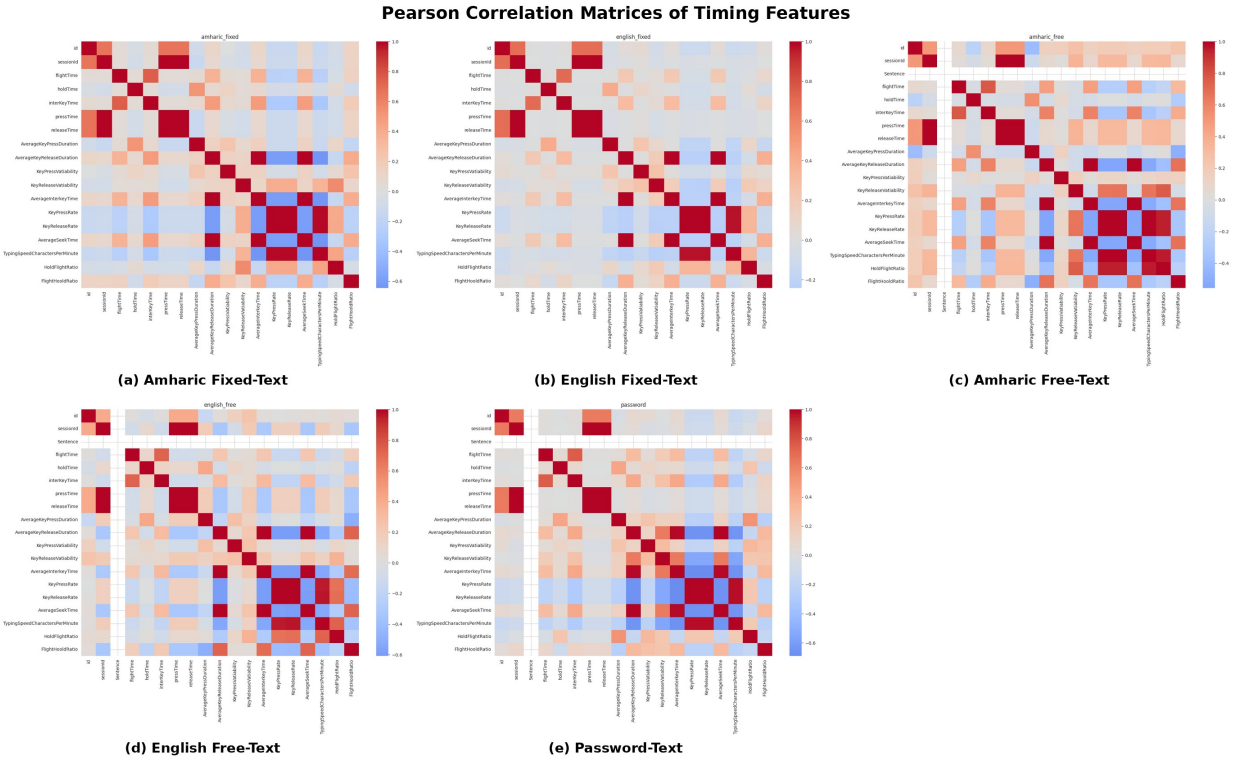


Figure 7. Pearson correlation matrices of the numeric timing-feature set for each sub-corpus. Dark red/blue blocks indicate near-perfect positive/negative redundancy between engineered and raw features.

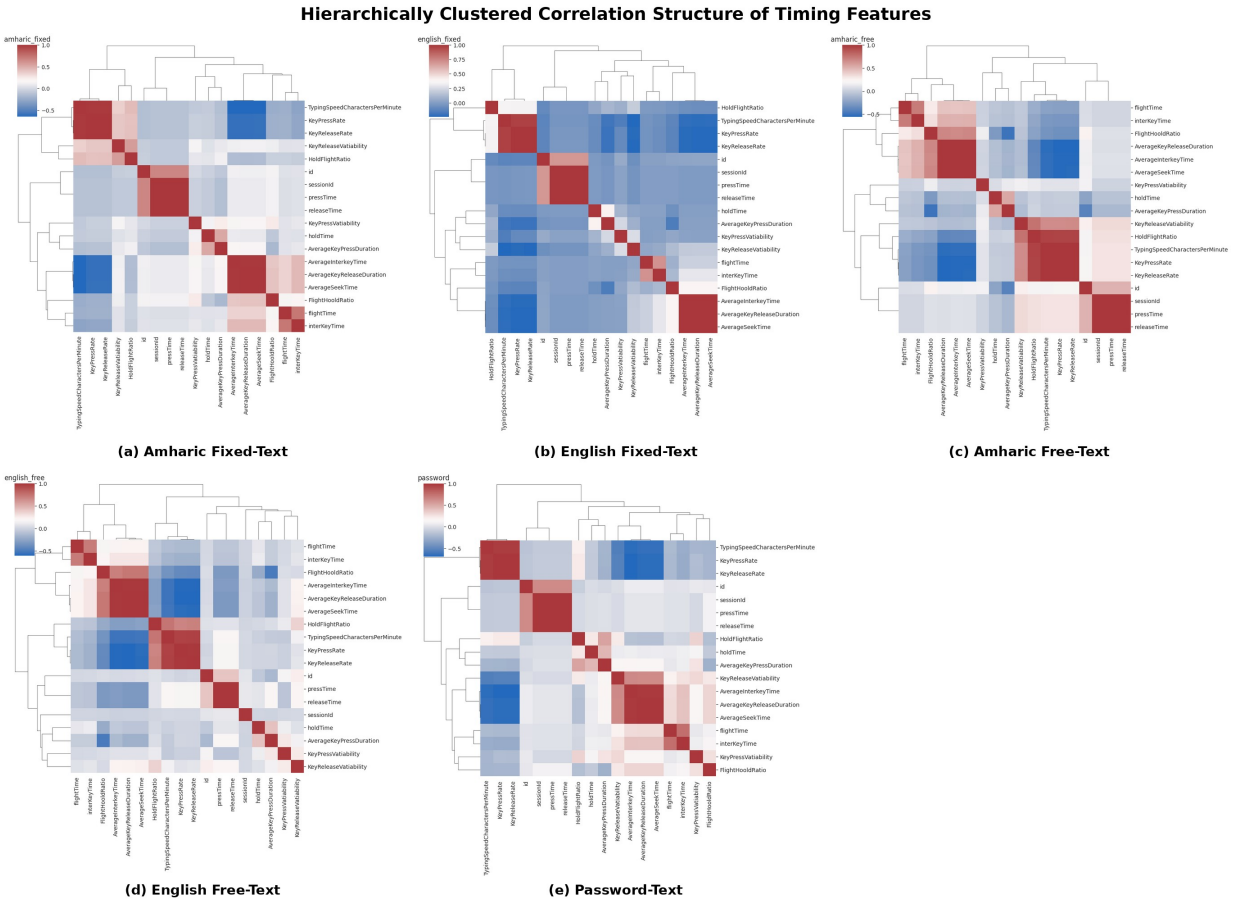


Figure 8. Hierarchically clustered correlation heatmaps (cluster maps), grouping algebraically related features and exposing an effective dimensionality substantially lower than the nominal feature count.

4.7 Per-user contribution imbalance

Table table5 summarizes the number of samples each user contributed to each sub-corpus. All sub-corpora are strongly right-skewed: median contributions (46–103 samples per user) are well below the means (139.7–283.1), while maxima (722–5,145) are roughly 13 to 65 times the corresponding medians. English free-text is the most imbalanced, with a coefficient of variation (SD/mean) of 2.43, suggesting that a small number of highly active users provide a large share of the training signal. Figure figure9 illustrates this pattern. Although common in behavioral-biometric field data, this imbalance is seldom acknowledged. It poses specific challenges for closed-set user identification, where over-represented users are easier to learn, while macro-averaged accuracy declines for under-represented users. It also affects open-set user verification, where over-represented users dominate EER estimates unless user- or session-level weighting is applied. Recent benchmarks show that reported EERs depend on per-subject versus global thresholding (Gonzálezxet al., 2025), underscoring the need to make these choices explicit.

Table 5. Distribution of samples per user within each sub-corpus

Sub-corpus	Users (n)	Mean	SD	Min	Median	Max	CV (SD/mean)
amharic_fixed	95	196.7	298.6	12	49.0	1,526	1.52
english_fixed	183	283.1	438.0	9	103.0	2,562	1.55
amharic_free	34	139.7	199.9	6	53.5	722	1.43
english_free	75	272.2	662.2	4	79.0	5,145	2.43
password	121	159.4	317.1	13	46.0	1,673	1.99

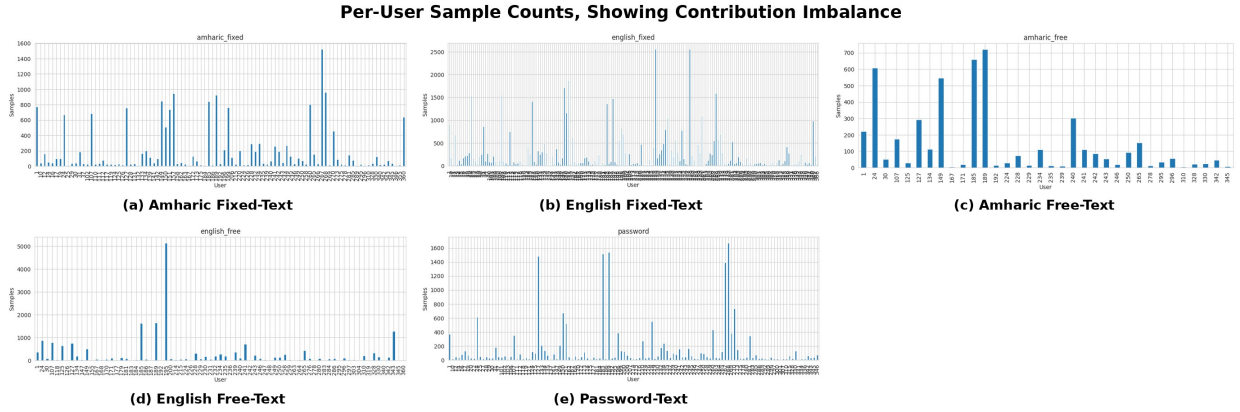


Figure 9. Per-user sample counts by sub-corpus (users sorted by identifier), illustrating substantial contribution imbalance in every modality.

4.8 User separability in reduced feature space

To assess user separability in the standardized timing-feature space before supervised training, we projected each sub-corpus using PCA (linear; Figure figure10), t-SNE (nonlinear and focused on local structure; Figure figure11), and UMAP (nonlinear and designed to balance local and global structure; Figure figure12). Points are colored by participant identity. Across all three methods and all five sub-corpora, participants overlap substantially. No sub-corpus shows compact, well-separated user clusters that would suggest near-perfect classification from this feature space. The PCA projections form a single, diffuse, largely unimodal cloud, suggesting that the first two principal components capture task-level variation (e.g., overall speed) rather than user-discriminative structure. These results are consistent with, and may help explain, the baseline classifiers' previously obtained rank-1 accuracy of 40–60% and EER of 10–20% on this cohort. If three complementary projections of the 18–19-dimensional feature space reveal no clear user structure, near-perfect discrimination should not be expected from downstream classifiers using the same features without enrichment (e.g., sequence- or digraph-level representations) or session- and user-level adaptation. Because two-dimensional embeddings distort high-dimensional neighborhoods (Charix& Pachter, 2023; Wangxet al., 2021), this evidence is qualitative and should be complemented by quantitative separability measures in future work.

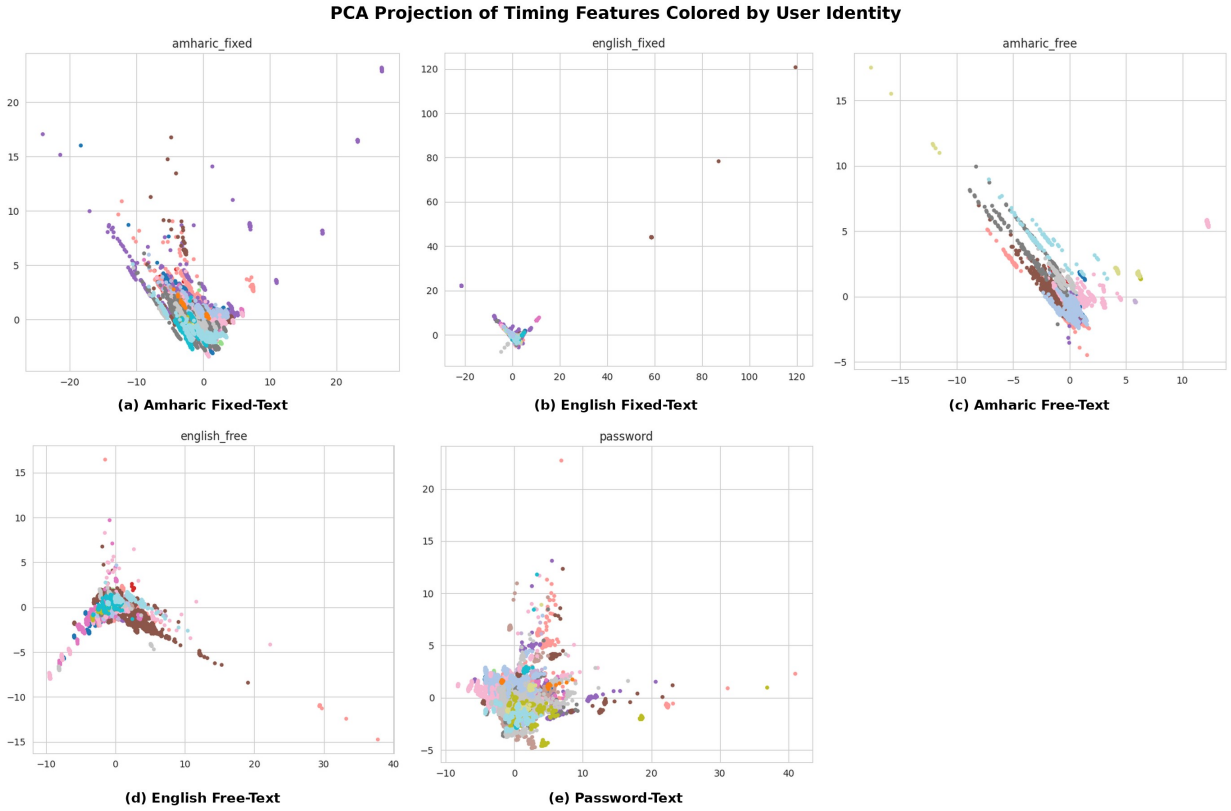


Figure 10. PCA projections (first two principal components) of standardized timing features, colored by participant identity. Substantial overlap is visible in every sub-corpus.

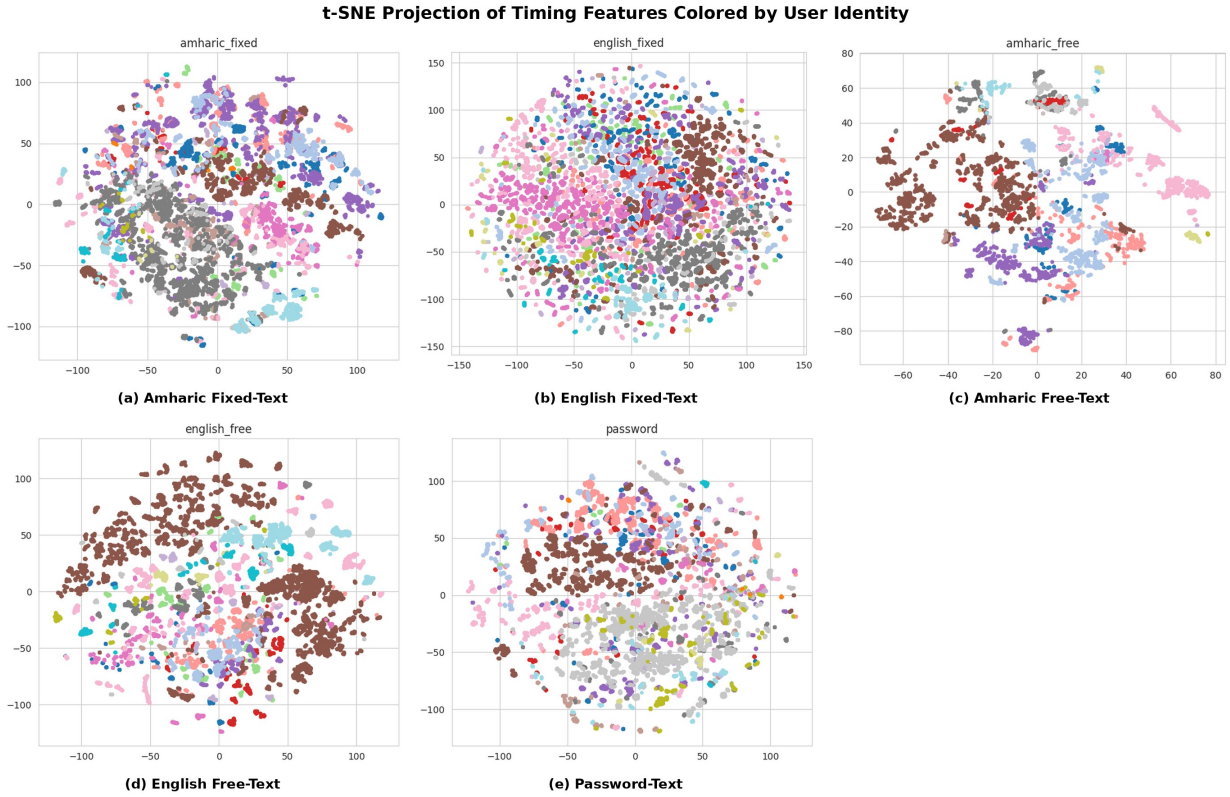


Figure 11. t-SNE projections of standardized timing features, colored by participant identity (perplexity default, random state 42). Local grouping is visible, but inter-user mixing remains substantial.

Figure 12. UMAP projections of standardized timing features, colored by participant identity (default parameters, random state 42), shown for comparison against t-SNE.

4.9 Fixed-text versus free-text comparison

Figure figure13 compares typing-speed distributions across the pooled fixed-text and free-text sub-corpora, focusing on the two text-constraint modes. As expected, free-text entry has a higher median speed and greater variability, since typing alternates between fluent production and pauses for composition. This pattern aligns with the heavy upper tail in the flight-time distributions observed in Section subsection4.4, while fixed-text entry, which uses short prompts typed repeatedly, produces a tighter speed distribution.

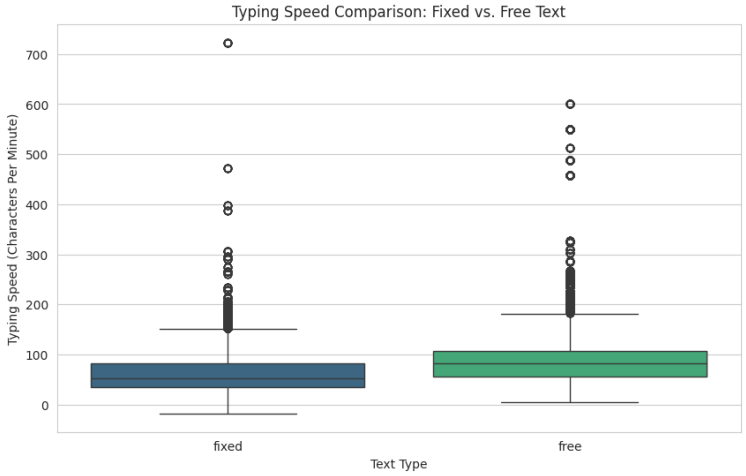


Figure 13. Typing-speed comparison between pooled fixed-text and pooled free-text sub-corpora.

4.10 Keystroke-content profiling

Finally, Figures figure14 and figure15 profile raw key-level content across each sub-corpus. Key-type distributions (Figure figure14) are dominated by alphabetic and character keys in all sub-corpora, while the password sub-corpus shifts markedly toward digit and symbol keys, consistent with common password-composition requirements. The top-30 key-frequency distributions (Figure figure15) show that the Amharic sub-corpora use a broad range of *fidel* characters (correctly rendered with Ethiopic fonts), whereas the English sub-corpora concentrate most of their frequency mass on a few high-frequency Latin characters. This difference in symbol-space size and frequency structure is an important explanatory factor. We propose that future models capture it explicitly, for example, through key-identity embeddings of the kind used alongside timing features in recent deep keystroke models (Acienxet al., 2022; Stragapede, Vera-Rodriguez,xet al., 2024).

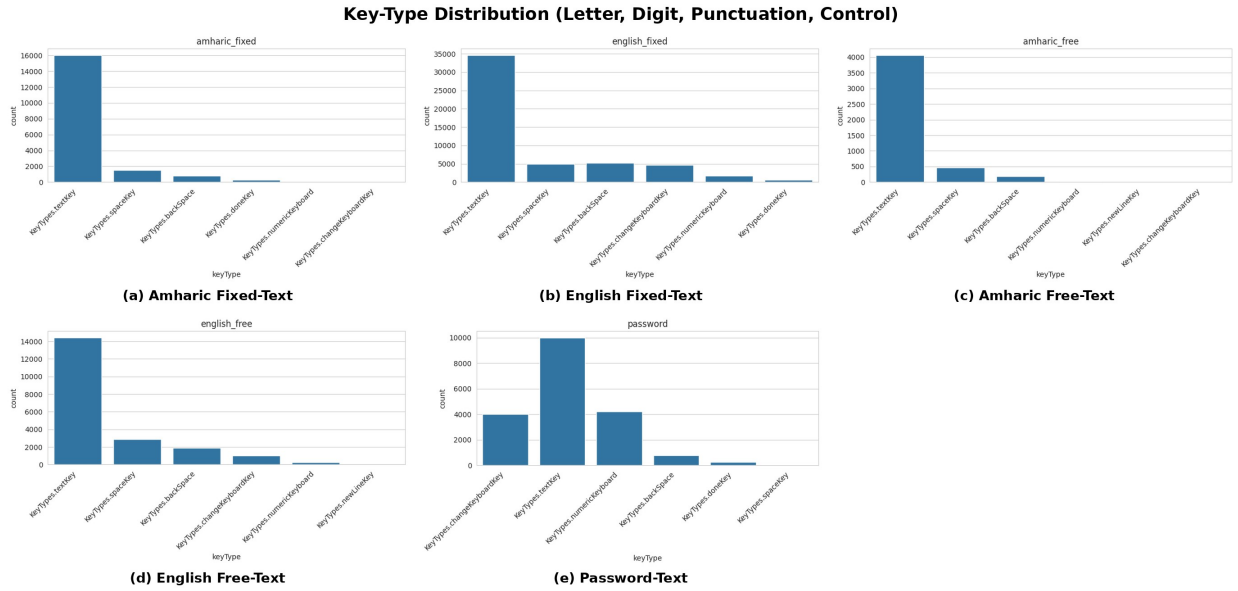


Figure 14. Key-type distribution (letter, digit, punctuation, control, etc.) by sub-corpus.

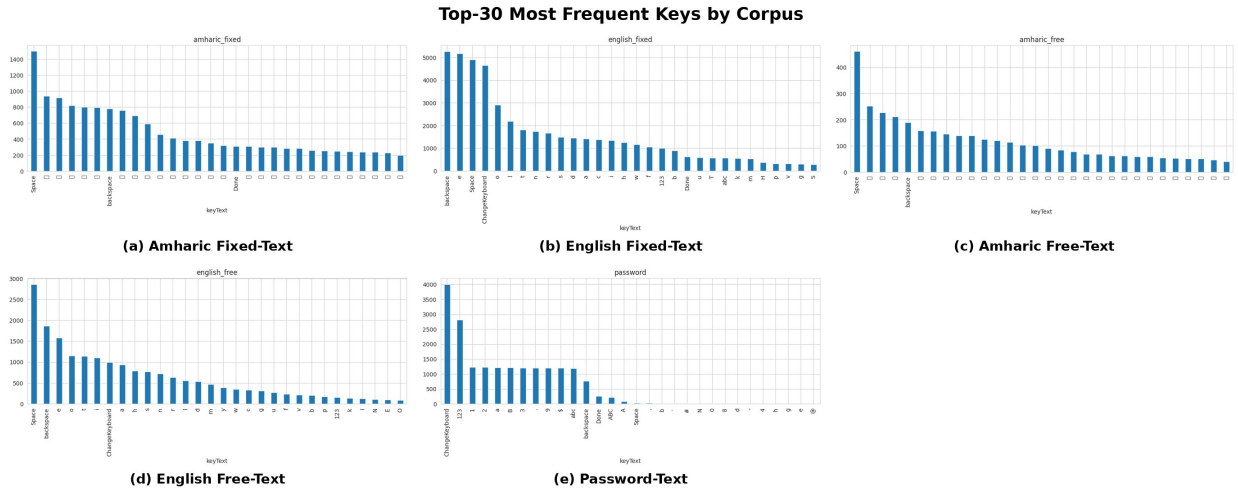


Figure 15. Top 30 most frequent keys by sub-corpus, illustrating Amharic's larger, script-specific symbol space relative to English.

5 Discussion

5.1 Explaining the baseline performance ceiling

Taken together, four EDA findings offer a coherent, testable explanation for the baseline results on this corpus: 40–60% accuracy and 10–20% EER. First, the extreme right skew and outlier-driven variance in flight and inter-key times (Sections subsection4.4–subsection4.5) mean that naive mean- and variance-based features are dominated by rare events and do not represent a user's typical rhythm well. Second, algebraic redundancy among engineered features (Section subsection4.6) indicates that the feature space's effective dimensionality is far below its nominal 18–19

columns, leaving models with fewer independent discriminative features. Third, the heavy-tailed distribution of per-user sample counts (Section subsection4.7) causes empirical-risk-minimization learners to fit a few prolific users at the expense of macro-averaged performance. Fourth, and most directly related to the classification ceiling, substantial overlap among users in PCA, t-SNE, and UMAP (Section subsection4.8) suggests that the row-level timing representation analyzed here is not sufficiently user-discriminative on its own. These factors interact with corpus scale: with 34–183 users per sub-corpus and few samples for many users, the current data fall into a regime where data-hungry deep models may underperform simpler statistical matchers (Wahabx& Hou, 2023), unlike the large Aalto-based benchmarks on which deep models excel (Acienxet al., 2022; Stragapedexet al., 2025). Language and script effects may add further variability, given the cross-language degradation reported for bilingual typists (Wahabxet al., 2022).

5.2 Recommendations for modeling

Building on this diagnosis and the literature reviewed in Section section2, we identify six priorities for future modeling with this corpus. (1) Move beyond row-level and session-aggregate features toward sequence-level representations, such as per-session digraph and trigraph sequences processed by recurrent, convolutional, or attention-based encoders (Acienxet al., 2022; Gonzálezxet al., 2025; Stragapede, Delgado-Santos,xet al., 2024). The redundancy analysis suggests that current aggregate features provide little independent information. (2) Replace random row-level splits with session- and user-disjoint training, validation, and test partitions. Since rows from the same session share recomputed aggregates, random splits would leak these features across partitions. Dependence between training and test samples is also a well-documented source of inflated performance (Kapoorx& Narayanan, 2023). Standardized protocols such as KVC offer a model for organizing fair comparisons (Stragapede, Vera-Rodriguez,xet al., 2024). (3) Use outlier-robust statistics, such as median- and MAD-based normalization, or model long pauses explicitly as separate features, since outliers contribute substantially to variance (Section subsection4.5). Treat imputation as an explicit modeling decision, because missingness is structured rather than random (Section subsection4.2). (4) Use methods suited to the small and uneven number of samples per user, such as metric-learning embeddings trained with contrastive, triplet, or Siamese objectives (Acienxet al., 2022; Lisxet al., 2023), as well as instance-based statistical matchers such as ITAD (Ayottexet al., 2020), which can match or outperform deep models on small datasets (Wahabx& Hou, 2023). (5) Encode key identity and script information explicitly (Section subsection4.10), rather than pooling Amharic and English rows in a single feature space, given their different symbol spaces and timing distributions, as well as the cross-language degradation reported for bilingual typists (Wahabxet al., 2022). When privacy is a concern, timing-only features offer a less invasive alternative (Stragapede, Vera-Rodriguez,xet al., 2024). (6) Report error rates using explicit threshold policies (per-user versus global) and disaggregate them by gender and skill level, because aggregate EERs can conceal threshold effects (Gonzálezxet al., 2025) and demographic differentials (Drozdowskixet al., 2020; Stragapedexet al., 2025).

5.3 Implications for script-inclusive behavioral biometrics

Beyond these modeling recommendations, the corpus's broader contribution is evidential: it demonstrates that a complete, standard keystroke-dynamics EDA and baseline workflow can be

applied to a language written in the Ethiopic script. The overall findings—skewed timing distributions, redundant engineered features, imbalanced user contributions, and overlapping user manifolds—are consistent with challenges widely reported for free-text keystroke data in Latin-script corpora, including high intra-subject variability and sensitivity to text content (Acienxet al., 2022; Shadmanxet al., 2025), although the performance ceiling observed here is lower. The English sub-corpora provide an internal reference processed with identical features, but comparisons with external Latin-script corpora remain qualitative because their features and protocols differ. These results suggest that the methodological toolkit developed for Latin-script keystroke dynamics is a reasonable starting point for non-Latin scripts. However, they also show that dataset-specific diagnostics, such as those presented here, are needed before generalizing, particularly given evidence that language and alphabet affect keystroke templates (Kang, 2015; Wahabxet al., 2022) and the broader under-representation of Ethiopian languages in language technology (Tonjaxet al., 2024; Yimamxet al., 2021).

6 Limitations and Threats to Validity

Several limitations qualify these conclusions. First, participants self-reported their gender and skill levels, which we did not independently verify. Thus, the subgroup differences discussed in Section subsection4.3 may reflect this sample rather than the broader population of Amharic and English typists. Second, although overlapping identifier ranges suggest the five sub-corpora came from a largely shared pool, no explicit linkage table was available. Therefore, treat the comparisons in Section subsection4.9 as suggestive, not as a fully controlled within-subject design. Third, the negative and otherwise implausible timing values reported in Section subsection4.5 point to logging artifacts that should be cleaned before high-stakes model training. However, if flight time is defined as the interval from releasing one key to pressing the next, some negative values may reflect genuine key overlap (rollover), not clock skew. We should distinguish these cases during cleaning. Fourth, the dimensionality-reduction analysis in Section subsection4.8 is graphical and qualitative. We did not compute quantitative cluster-separation measures (e.g., silhouette scores) to avoid implying more precision than the EDA supports, and two-dimensional embeddings are known to distort high-dimensional structure (Charix& Pachter, 2023). Fifth, missing values were replaced with zeros or column means before the correlation and projection analyses (Section subsection3.4), which may attenuate or distort some correlations. Sixth, the data schema does not describe the Amharic text-entry method (e.g., the specific transliteration input method or keyboard layout), limiting interpretation of Amharic timing features. Finally, this paper presents EDA and baseline characterization, not new classifier results. The 40–60% accuracy and 10–20% EER values cited throughout come from earlier baseline experiments on this cohort that are not reported in detail here. Comparisons with error rates published for other corpora are also not like-for-like because datasets, protocols, and metrics differ.

7 Conclusion and Future Work

We developed and characterized a multimodal, bilingual keystroke dynamics corpus covering fixed-text, free-text, and password modalities in Amharic and English. Data were collected from a shared participant pool, with 34–183 participants per sub-corpus, and comprise 114,955 labeled keystroke events. To the best of our knowledge, this is the first comprehensive keystroke dynamics

corpus to include an Ethiopic-script language, addressing a gap in behavioral biometrics research dominated by Latin-script and English-language data. Our EDA pipeline examines corpus structure, data quality, demographics, univariate and bivariate timing distributions, feature redundancy, per-user imbalance, and separability using PCA, t-SNE, and UMAP. It generates reproducible diagnostics that help explain baseline classifier performance and inform the modeling recommendations in Section subsection5.2. Future work should (i) develop and evaluate sequence-level, session-aware, and few-shot models using session- and user-disjoint protocols, as outlined in Section subsection5.2; (ii) extend the corpus with additional behavioral modalities (touch, mouse, and gaze), following the original collection's multimodal fusion rationale; (iii) expand the participant pool and improve gender and proficiency balance; (iv) document the Amharic input method and add quantitative separability and fairness analyses; and (v) broaden script coverage to include other under-represented writing systems, creating more representative behavioral-biometric benchmarks.

Data and Code Availability

The complete keystroke dynamics dataset supporting the findings of this study is publicly available at https://github.com/animawkerie/BahriApp-Dataset. The exploratory data analysis workflow is implemented in a standalone Jupyter notebook, which is available from the corresponding author upon reasonable request, subject to participant confidentiality and institutional data-sharing rules.

Declarations

Conflict of interest: Authors have no conflicts of interest to declare.

Ethical considerations: The study was conducted as part of the corresponding author's doctoral research and received approval from the Addis Ababa Science and Technology University Institutional Ethical Approval Board (AASTU-IEAB) under form number RTT/RCD/OF/0028. All data-handling procedures followed the approved protocol and Ethiopia's Personal Data Protection Proclamation No. 1321/2024. Participation was voluntary and limited to adults, so guardian consent was not required. Before data capture, the mobile data collection application informed each participant of the study's purpose, the behavioral data to be recorded, the plan to publicly release anonymized data, and their right to withdraw at any time. Participants then consented to take part and share their data through an explicit in-app confirmation.

The protocol permitted anonymous collection only. The released files stored no names, contact details, national or institutional identification numbers, or facial images. Instead, each participant is represented by a random numeric identifier, which is the only key linking records across sessions and modalities. The only personal attributes retained are self-reported gender and typing-skill level. We captured data using version 1.0.0 of the BahriApp custom-developed mobile application. Because behavioral signals can support re-identification, the released files contain no direct identifiers, and a Data Usage Agreement governs reuse and prohibits any attempt to re-identify participants.

References

Abuhamad, M., Abusnaina, A., Nyang, D., & Mohaisen, D. (2021). Sensor-based continuous authentication of smartphones' users using behavioral biometrics:

A contemporary survey. *IEEE Internet of Things Journal*, *8*(1), 65–84. doi: https://doi.org/10.1109/JIOT.2020.3020076

Acien, A., Morales, A., Monaco, J. V., Vera-Rodriguez, R., & Fierrez, J. (2022). TypeNet: Deep learning keystroke biometrics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *4*(1), 57–70. doi: https://doi.org/10.1109/tbiom.2021.3112540

Ayotte, B., Banavar, M., Hou, D., & Schuckers, S. (2020). Fast free-text authentication via instance-based keystroke dynamics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *2*(4), 377–387. doi: https://doi.org/10.1109/tbiom.2020.3003988

Belay, B., Habtegebrial, T., Meshesha, M., Liwicki, M., Belay, G., & Stricker, D. (2020). Amharic OCR: An end-to-end learning. *Applied Sciences*, *10*(3), 1117. doi: https://doi.org/10.3390/app10031117

Chari, T., & Pachter, L. (2023). The specious art of single-cell genomics. *PLOS Computational Biology*, *19*(8), e1011288. doi: https://doi.org/10.1371/journal.pcbi.1011288

Delgado-Santos, P., Stragapede, G., Tolosana, R., Guest, R., Deravi, F., & Vera-Rodriguez, R. (2022). A survey of privacy vulnerabilities of mobile device sensors. *ACM Computing Surveys*, *54*(11s), 1–30. doi: https://doi.org/10.1145/3510579

Dhakal, V., Feit, A. M., Kristensson, P. O., & Oulasvirta, A. (2018). Observations on typing from 136 million keystrokes. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–12). ACM. doi: https://doi.org/10.1145/3173574.3174220

Dias, T., Vitorino, J., Maia, E., Sousa, O., & Praça, I. (2023). KeyRecs: A keystroke dynamics and typing pattern recognition dataset. *Data in Brief*, *50*, 109509. doi: https://doi.org/10.1016/j.dib.2023.109509

Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., & Busch, C. (2020). Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, *1*(2), 89–103. doi: https://doi.org/10.1109/TTS.2020.2992344

Giot, R., El-Abed, M., & Rosenberger, C. (2009). GREYC keystroke: A benchmark for keystroke dynamics biometric systems. In *2009 IEEE 3rd international conference on biometrics: Theory, applications, and systems* (pp. 1–6). IEEE. doi: https://doi.org/10.1109/btas.2009.5339051

González, N., Calot, E. P., Ierache, J. S., & Hasperué, W. (2022). Towards liveness detection in keystroke dynamics: Revealing synthetic forgeries. *Systems and Soft Computing*, *4*, 200037. doi: https://doi.org/10.1016/j.sasc.2022.200037

González, N., Stragapede, G., Vera-Rodriguez, R., & Tolosana, R. (2025). Type2Branch: Keystroke biometrics based on a dual-branch architecture with attention mechanisms and set2set loss. *IEEE Transactions on Information Forensics and Security*, *20*, 5859–5871. doi: https://doi.org/10.1109/TIFS.2025.3574992

Indiana University. (n.d.). *Amharic [Language information portal]*. `https://celt.indiana.edu/portal/Amharic`. (Retrieved October 28, 2026)

Kang, P. (2015). The effects of different alphabets on free text keystroke authentication: A case study on the Korean–English users. *Journal of Systems and Software*, *102*, 1–11. doi: https://doi.org/10.1016/j.jss.2014.12.017

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), 100804. doi: https://doi.org/10.1016/j.patter.2023.100804

Killourhy, K. S., & Maxion, R. A. (2009). Comparing anomaly-detection algorithms for keystroke dynamics. In *2009 IEEE/IFIP international conference on dependable systems & networks*

(pp. 125–134). IEEE. doi: https://doi.org/10.1109/dsn.2009.5270346

Kim, J., & Kang, P. (2020). Freely typed keystroke dynamics-based user authentication for mobile devices based on heterogeneous features. *Pattern Recognition*, *108*, 107556. doi: https://doi.org/10.1016/j.patcog.2020.107556

Lis, K., Niewiadomska-Szynkiewicz, E., & Dziewulska, K. (2023). Siamese neural network for keystroke dynamics-based authentication on partial passwords. *Sensors*, *23*(15), 6685. doi: https://doi.org/10.3390/s23156685

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, *3*(29), 861. Retrieved from http://dx.doi.org/10.21105/joss.00861 doi: https://doi.org/10.21105/joss.00861

Microsoft. (n.d.). *Amharic IME.* Microsoft Learn. https://learn.microsoft.com/globalization/input/amharic-ime. (Retrieved October 28, 2026)

Shadman, R., Wahab, A. A., Manno, M., Lukaszewski, M., Hou, D., & Hussain, F. (2025). Keystroke dynamics: Concepts, techniques, and applications. *ACM Computing Surveys*, *57*(11), 1–35. doi: https://doi.org/10.1145/3733103

Sitová, Z., Šeděnka, J., Yang, Q., Peng, G., Zhou, G., Gasti, P., & Balagani, K. S. (2016). HMOG: New behavioral biometric features for continuous authentication of smartphone users. *IEEE Transactions on Information Forensics and Security*, *11*(5), 877–892. doi: https://doi.org/10.1109/tifs.2015.2506542

Stragapede, G., Delgado-Santos, P., Tolosana, R., Vera-Rodriguez, R., Guest, R., & Morales, A. (2023). Mobile keystroke biometrics using transformers. In *2023 IEEE 17th international conference on automatic face and gesture recognition (FG)* (pp. 1–6). IEEE. doi: https://doi.org/10.1109/FG57933.2023.10042710

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. doi: https://doi.org/10.1007/s00521-024-10140-2

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

Stragapede, G., Vera-Rodriguez, R., Tolosana, R., Morales, A., Damer, N., Fierrez, J., & Ortega-Garcia, J. (2024). Keystroke verification challenge (KVC): Biometric and fairness benchmark evaluation. *IEEE Access*, *12*, 1102–1116. doi: https://doi.org/10.1109/ACCESS.2023.3345452

Stragapede, G., Vera-Rodriguez, R., Tolosana, R., Morales, A., DeAndres-Tame, I., Damer, N., ... others (2025). KVC-ongoing: Keystroke verification challenge. *Pattern Recognition*, *161*, 111287. doi: https://doi.org/10.1016/j.patcog.2024.111287

Stragapede, G., Vera-Rodriguez, R., Tolosana, R., Morales, A., Fierrez, J., Ortega-Garcia, J., ... others (2022). IJCB 2022 mobile behavioral biometrics competition (MobileB2C). In *2022 IEEE international joint conference on biometrics (IJCB)* (pp. 1–7). IEEE. doi: https://doi.org/10.1109/IJCB54206.2022.10007985

Stylios, I., Kokolakis, S., Thanou, O., & Chatzis, S. (2021). Behavioral biometrics & continuous user authentication on mobile devices: A survey. *Information Fusion*, *66*, 76–99. doi: https://doi.org/10.1016/j.inffus.2020.08.021

Sun, Y., Ceker, H., & Upadhyaya, S. (2016). Shared keystroke dataset for continuous authentication. In *2016 IEEE international workshop on information forensics and security (WIFS)*

(pp. 1–6). IEEE. doi: https://doi.org/10.1109/WIFS.2016.7823894

Tonja, A. L., Azime, I. A., Belay, T. D., Yigezu, M. G., Mehamed, M. A. A., Ayele, A. A., … Yimam, S. M. (2024). EthioLLM: Multilingual large language models for Ethiopian languages with task evaluation. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 6341–6352). ELRA and ICCL. doi: https://doi.org/10.63317/3xseji2kaeg2

Tsimperidis, I., Yucel, C., & Katos, V. (2021). Age and gender as cyber attribution features in keystroke dynamic-based user classification processes. *Electronics*, *10*(7), 835. doi: https://doi.org/10.3390/electronics10070835

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

Wahab, A. A., & Hou, D. (2023). When simple statistical algorithms outperform deep learning: A case of keystroke dynamics. In *Proceedings of the 12th international conference on pattern recognition applications and methods (ICPRAM)* (Vol. 1, pp. 363–370). SCITEPRESS. doi: https://doi.org/10.5220/0011684100003411

Wahab, A. A., Hou, D., Banavar, M., Schuckers, S., Eaton, K., Baldwin, J., & Wright, R. (2022). Shared multi-keyboard and bilingual datasets to support keystroke dynamics research. In *Proceedings of the twelfth ACM conference on data and application security and privacy* (pp. 236–241). ACM. doi: https://doi.org/10.1145/3508398.3511516

Wang, Y., Huang, H., Rudin, C., & Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. *Journal of Machine Learning Research*, *22*(201), 1–73.

Yimam, S. M., Ayele, A. A., Venkatesh, G., Gashaw, I., & Biemann, C. (2021). Introducing various semantic models for Amharic: Experimentation and evaluation with multiple tasks and datasets. *Future Internet*, *13*(11), 275. doi: https://doi.org/10.3390/fi13110275