DOI: https://doi.org/10.20535/kpisn.2025.4.344357
UDC 004.032.26:004.93

M.P. Havrylovych
Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine
*Corresponding author: mariia.havrylovych@gmail.com

ARCHITECTURE OF HYBRID CNN-TRANSFORMER WITH MASKED TIME SERIES AUTO-CODING FOR BEHAVIOURAL BIOMETRICS ON MOBILE DEVICES

Background. Continuous behavioural authentication (keystroke dynamics, touch/swipe, motion sensors) verifies identity without extra actions. However, models degrade under device, session and activity shifts, are sensitive to noise and often require significant labelling. As passwordless logins spread, demand rises for post-login risk control and for models that are robust, compute-efficient and stable in real-world conditions.

Objective. The paper aims to develop and empirically study a compact CNN-Transformer hybrid with lightweight self-supervised masked time-series autoencoding (MAE-style) for mobile behavioural biometrics on the HMOG and WISDM datasets.

Methods. A 1D-CNN front end extracts local cues from smartphone motion signals, while a Transformer encoder captures longer-range dependencies. We use masked reconstruction on unlabelled HMOG sessions for self-supervised pretraining under a limited computational budget and then fine-tune the same hybrid architecture for user identification. We evaluate three hybrid variants on HMOG (trained from scratch, with masked pretraining, and with masked pretraining plus CORAL domain adaptation) and three models on WISDM (a Transformer baseline, a hybrid trained from scratch and a hybrid initialised from the HMOG-pretrained weights). Performance is measured using user-level mean and median Equal Error Rate (EER) and AUC at the individual user level.

Results. On HMOG, the hybrid model trained from scratch achieves the best user-level metrics (EER 21.51 % mean, 18.63 % median; AUC 0.854 mean, 0.905 median), while the lightweight MAE and CORAL variants do not yet surpass this baseline. On WISDM, the hybrid model substantially outperforms a pure Transformer baseline (EER 9.41 % vs 51.25 % mean; AUC 0.902 vs 0.488 mean), and cross-dataset initialisation from the HMOG MAE-pretrained weights provides an additional improvement (EER 8.42 % mean, 2.07 % median; AUC 0.907 mean, 0.959 median).

Conclusions. The results indicate that a compact CNN-Transformer hybrid is effective for sensor-based mobile behavioural biometrics and that even lightweight masked pretraining can be helpful for cross-dataset transfer. At the same time, the benefits of MAE and CORAL on HMOG depend strongly on the pretraining budget and masking configuration, suggesting that further tuning is needed to fully exploit self-supervised pretraining in this setting.

Keywords: behavioural biometrics; continuous authentication; smartphone sensors; CNN-Transformer hybrid; masked autoencoding; self-supervised pretraining; domain adaptation.

Introduction

The widespread use of smartphones and wearables has turned them into primary access points for services, including systems that explicitly explore behavioural biometrics on everyday activities [1] and continuous sensing on smartphones [2, 3], which in turn creates stricter requirements for the underlying information security mechanisms. Traditional one-shot authentication methods such as passwords, PIN codes and fingerprint scans verify the user only at login time. Once a device is unlocked, anyone who physically gains access to it can continue to work under the legitimate user's identity. This is particularly critical when smartphones are used to access financial services, corporate resources and personal communications, as demonstrated both in general-purpose smartphone biometrics [1, 2, 3] and in our earlier work on continuous authentication for security-critical services [4].

Continuous behavioural authentication offers an alternative paradigm: the user's identity is verified

Пропозиція для цитування цієї статті: М.П. Гаврилович, «Архітектура гібридного CNN-transformer з маскованим автокодуванням часових рядів для поведінкової біометрії на мобільних пристроях», *Наукові вісті КПІ*, № 4, с. 55–62, 2025. doi: https://doi.org/10.20535/kpisn.2025.4.344357

Offer a citation for this article: M.P. Havrylovych, "Architecture of hybrid CNN-transformer with masked time series auto-coding for behavioural biometrics on mobile devices", *KPI Science News*, No. 4, pp. 55–62, 2025. doi: https://doi.org/10.20535/kpisn.2025.4.344357

in the background throughout device usage, based on behavioural signals [1, 2, 3]. These signals include keystroke dynamics on the virtual keyboard, which have been extensively reviewed for both fixed-text and free-text scenarios [8, 9], deep keystroke models on desktop and mobile platforms [10, 11], as well as touch/swipe patterns and inertial sensor data such as accelerometer and gyroscope signals that underpin smartphone and smartwatch biometrics [1, 2, 3]. Together, they form a behavioural "fingerprint" that can be used to distinguish one user from others without requiring explicit re-authentication. This class of methods is closely related to behavioural biometrics and continuous authentication frameworks used for post-login risk control in high-stakes applications [1, 4].

However, building robust behavioural biometric models is challenging. Unlike static biometrics, behavioural patterns are highly context-dependent. They vary with posture (sitting, standing, walking), activity, device model and UI layout, and can also drift over time. Sensor data is noisy and often contains missing values. Changes in hardware, operating system version or user habits can cause domain shifts that degrade the performance of models trained on earlier data. Collecting large labelled datasets per user is expensive and often impractical, especially at scale, a limitation repeatedly highlighted in smartphone and sensor-based continuous authentication studies [1, 2, 3] and confirmed in our own experiments on motion-based verification and wearable sensing [4].

Recent advances in deep learning, particularly convolutional and recurrent architectures in continuous authentication [4, 6] and Transformer-based models for keystroke and time-series data [10, 11, 13, 14], have significantly improved the state of the art in signal and sequence modelling. CNNs are effective at capturing local patterns and invariances, while Transformers use self-attention to model long-range dependencies. In parallel, self-supervised learning methods such as masked autoencoders (MAE) have demonstrated that useful representations can be learned from large unlabelled datasets by reconstructing masked parts of the input [13, 14].

Despite these advances, many mobile behavioural biometric systems still rely on purely convolutional or recurrent architectures [1, 2, 11] or on traditional keystroke pipelines surveyed in [8, 9], are trained from scratch on relatively small labelled datasets and only partially address domain shifts between sessions and conditions. There is still a need for models that can exploit unlabelled behavioural data, maintain robustness under cross-session and cross-condition scenarios and remain efficient enough for deployment on mobile devices.

This work addresses these challenges by exploring a hybrid CNN-Transformer architecture with masked time-series autoencoding for mobile behavioural biometrics. The proposed model targets continuous authentication scenarios, where decisions about user identity must be made based on short sliding windows of behavioural data, similar to the window-based protocols used in HMOG, WISDM and related continuous authentication work [2, 3, 12]. The architecture is designed to leverage unlabelled sessions for lightweight self-supervised pretraining and to support efficient inference on mobile devices while remaining robust to domain shifts, building on ideas from prior deep continuous authentication systems [4, 6], Transformer-based keystroke and time-series models [10, 11, 14] and masked autoencoding for temporal data [13].

Problem Statement

Let $U = \{u\_1, ..., u\_\mathrm{K}\}$ be a set of users. For each user u_k, we have a collection of interaction sessions recorded from a smartphone. Each session consists of one or more time-series channels derived from sensors (e.g., accelerometer, gyroscope) and/or interaction events (such as touch coordinates). A session can be represented as a sequence $x\hat{}(i) = \{x\_t\hat{}(i)\}\_t$, where $x\_t\hat{}(i) \in R\hat{}C$, is the feature vector at time t and C is the number of channels.

For continuous authentication, the data stream is segmented into overlapping windows of fixed length T, producing fragments $X\_j \in R\hat{}(T\times C)$, each labelled with the corresponding user ID y_j in $\{1, ..., K\}$. The primary task considered in this work is user identification: given a window X, predict the user label y. Formally, we seek a model f_θ: $R\hat{}(T\times C)$ to $\{1, ..., K\}$ that maps each window to a distribution over user classes and minimises identification and verification errors under realistic cross-session and cross-condition settings.

In this work, we focus on Equal Error Rate (EER) and the area under the ROC curve (AUC), computed at the user level, as the primary evaluation metrics for continuous authentication. For each user, we compute individual EER and AUC values and then aggregate them across users by taking the mean and median. Beyond these verification metrics, the model should also satisfy practical constraints such as robustness to domain shifts and efficient inference on resource-constrained mobile hardware, and remain compatible with model compression and quantisation in future deployments.

Presentation of the Main Research Results

1. Related Work

Research on behavioural biometrics for mobile devices can be broadly divided into three directions: keystroke dynamics on virtual keyboards, sensor-based activity and movement analysis, and multimodal fusion of interaction and sensor signals [1, 2, 3].

Keystroke-based authentication methods analyse timing information associated with keypress events: inter-keystroke intervals, key hold times and editing patterns. Early works focused on fixed-text scenarios, whereas more recent approaches consider free-text typing where the user enters arbitrary content [8, 9]. It has been shown that even in free-text conditions, typing patterns remain sufficiently stable to support user identification and verification when combined with appropriate sequence models, including modern deep learning architectures [9, 10, 11].

In sensor-based continuous authentication, datasets such as HMOG and WISDM have become standard benchmarks. HMOG provides inertial sensor readings, device orientation and touch events from smartphones in sitting and walking scenarios, enabling evaluation under motion-induced variability and fine-grained hand movement patterns [2]. WISDM includes accelerometer and gyroscope time series from smartphones and smartwatches collected during daily activities and has been used both for activity recognition [1] and for biometric identification when users are treated as classes, including in our earlier work on motion-based verification [4]. Deep learning models for these datasets range from convolutional and recurrent networks to architectures specifically designed for continuous smartphone authentication [1, 3], with our previous studies exploring autoencoder-based and hybrid transformer architectures for user verification on motion and wearable signals [4, 6].

Transformer-based models have recently been proposed in mobile behavioural biometrics to operate directly on sequences of interaction and sensor events [10, 11, 14]. On large-scale typing datasets, Transformer architectures have been shown to outperform classical recurrent networks by effectively modelling long sequences of interactions and their contextual dependencies; TypeFormer is one example of a mobile keystroke Transformer achieving state-of-the-art results [10]. Hybrid CNN-Transformer architectures and attention-based sequence models more generally have also been explored in time-series processing, where convolutional layers serve as a front end for local pattern extraction and sequence length reduction, while Transformer encoders model global dependencies [11, 14].

Self-supervised methods, in particular masked autoencoders, allow learning robust representations from unlabelled data by reconstructing masked parts of the input. For time series, TS-MAE demonstrates that masked reconstruction can significantly improve representation quality under limited labels and domain shifts [13], while broader surveys of Transformers in time series highlight both the strengths and open issues of such models for temporal data [14]. Domain adaptation techniques such as Deep CORAL achieve additional robustness by aligning feature distributions across domains [15]. In our previous work we have investigated autoencoder-based and recurrent models for biometric verification using motion and sensor signals, as well as hybrid Transformer-autoencoder architectures for continuous authentication on wearable devices, demonstrating competitive Equal Error Rates and flexibility across signal types [5, 6, 7]. The present work extends these ideas to a CNN-Transformer hybrid with self-supervised pretraining and domain adaptation tailored to mobile behavioural biometrics.

2. Proposed CNN-Transformer Architecture with Masked Autoencoding

2.1. Input Preprocessing

Raw time-series data from sensors and, where available, interaction logs are first normalised per channel by subtracting the mean and dividing by the standard deviation computed on the training set. Each session is then segmented into overlapping windows of fixed length T with a chosen stride. Windows with insufficient valid samples are discarded. Each window $X \in R\^(C \times T)$ is treated as a multi-channel time-series fragment. For implementation convenience, tensors can be rearranged to shape (C, T) for compatibility with one-dimensional convolutions. In the experiments reported in this paper, we focus on inertial sensor channels from HMOG and WISDM.

2.2. Convolutional Front End

The convolutional front end is a stack of 1D convolutional layers applied along the temporal dimension. Each layer consists of a convolution with a small kernel, batch normalisation and a non-linear activation such as GELU. Channel dimensionality is gradually increased across layers, allowing the network to capture increasingly complex local patterns while suppressing noise. A multi-scale design can be achieved by combining kernels of different sizes or using dilated convolutions. The result is a sequence of feature vectors of shape $T \times C\_out$ that summarise local behavioural patterns such as micro-movements and short-term dynamics in the motion signals.

2.3. Transformer Encoder

The CNN features are linearly projected into a d-dimensional space to form a sequence of embeddings. Positional encodings, such as sinusoidal or relative positional encodings, are added to represent temporal order. The resulting sequence is processed by a stack of Transformer encoder layers, each comprising multi-head self-attention and position-wise feed-forward networks with residual connections and layer normalisation. Self-attention allows the model to focus on the most informative events within the window and to capture long-range dependencies and interactions between channels. This is particularly important for behavioural biometrics, where discriminative patterns may be scattered across the window rather than localised.

2.4. Masked Time-Series Autoencoding

To exploit unlabelled sessions, a masked time-series autoencoding task is used for self-supervised pretraining. For each window, a binary mask over time steps is sampled, masking a fixed fraction of positions. The corresponding inputs are zeroed out, and the masked sequence is fed through the CNN front end and Transformer encoder. A reconstruction head maps the hidden representations back to the CNN feature space, and the mean squared error is computed between the reconstructed and original CNN features, but only on masked time steps.

This masked reconstruction objective encourages the model to infer missing local patterns from temporal context and to build representations that are robust to noise and missing data. Because no user labels are required, large volumes of unlabelled behavioural data can be used for pretraining. In practice, a lightweight pretraining regime is adopted: a compact model with modest dimensionality and a short window length is trained for a limited number of epochs and gradient steps, which is sufficient to provide a useful initialisation for subsequent supervised training.

2.5. Classification and Loss Functions

After the Transformer encoder, the sequence of hidden vectors is aggregated into a fixed-dimensional representation via global average pooling over time. The pooled vector is passed through a small multi-layer classification head consisting of layer normalisation, a hidden linear layer with non-linearity and an output linear layer mapping to K user classes.

In this work, the supervised loss is the standard cross-entropy loss. The architecture is compatible with angular-margin softmax losses and additional metric-learning losses such as triplet loss or center loss, as well as with domain adaptation regularizers such as CORAL, which we explicitly use in one of the HMOG variants. A more extensive exploration of alternative loss functions is left for future work.

3. Experimental Setup

3.1. Datasets

We consider two public datasets that are widely used in mobile behavioural biometrics and activity recognition.

The **HMOG** dataset provides multimodal recordings for continuous authentication, including accelerometer, gyroscope, magnetometer, device orientation and touch events from smartphones [2]. Users perform text-entry and other tasks in sitting and walking conditions, which enables evaluation under motion-induced variability. From HMOG we derive multimodal windows that may include multiple inertial sensor channels.

The **WISDM Smartphone and Smartwatch Activity and Biometrics** dataset contains accelerometer and gyroscope time series collected from multiple subjects during daily activities [1, 12]. While it is often used for activity recognition, we treat users as classes and extract fixed-length windows of motion data for biometric identification.

For both datasets, raw recordings are converted into fixed-length windows X in R^(T×C) with user labels. We keep the same windowing strategy across baselines and our model.

3.2. Evaluation Protocols

We use cross-session protocols in which training and testing data for each user come from different recording sessions. Where the dataset structure allows it, we additionally simulate cross-condition or cross-device scenarios by training and testing on disjoint subsets corresponding to different recording conditions or device types (for example, sitting versus walking conditions in HMOG).

Performance is reported in terms of user-level mean and median Equal Error Rate (EER) and the area under the ROC curve (AUC). For each user, we compute individual EER and AUC values and then aggregate them across users by taking the mean and median.

3.3. Model Variants and Baselines

To quantify the benefits of the proposed CNN-Transformer architecture, masked pretraining and domain adaptation, we evaluate a family of models on both datasets. Each model variant corresponds to a specific configuration in the codebase and is identified by a short experiment name.

On the **HMOG** dataset, we consider three variants:

– **HMOG_HYBRID_NO_MAE** – the proposed CNN-Transformer hybrid architecture trained from scratch in a purely supervised way, without masked pretraining or domain adaptation. This variant isolates the architectural contribution of the hybrid model.

– **HMOG_HYBRID_MAE_LIGHT** – the same hybrid architecture, but initialised using a lightweight masked autoencoding pretraining stage on HMOG. This experiment tests whether even modest self-supervised pretraining improves downstream identification and verification metrics.

– **HMOG_HYBRID_MAE_LIGHT_CORAL** – the hybrid model with lightweight MAE pretraining and an additional CORAL-based domain adaptation term that aligns feature distributions between two HMOG conditions (e.g., sitting vs walking) during supervised training. This variant is used to evaluate the impact of explicit domain adaptation on cross-condition performance.

On the **WISDM** dataset, we evaluate three analogous variants:

– **WISDM_TRANSFORMER** – a pure Transformer baseline trained on WISDM windows with no CNN front end.

– **WISDM_HYBRID_NO_MAE** – the CNN-Transformer hybrid architecture trained from scratch on WISDM without masked pretraining.

– **WISDM_HYBRID_FROM_HMOG_MAE** – the hybrid model initialised from the HMOG lightweight MAE-pretrained checkpoint and subsequently fine-tuned on WISDM. In this setting, no separate MAE pretraining is performed on WISDM; instead, HMOG serves as a source domain for cross-dataset pretraining.

Together, these experiments allow us to disentangle the effects of architecture (Transformer-only vs hybrid), masked pretraining (with vs without MAE) and domain adaptation (with vs without CORAL on HMOG), as well as to study the usefulness of cross-dataset pretraining when transferring from HMOG to WISDM.

3.4. Training Procedure and Ablation Studies

All models are trained using the same windowing strategy and train/validation splits within each dataset. The hybrid architecture is evaluated under different training regimes that correspond directly to the experiment list described above.

For hybrid models with masked pretraining, we adopt a lightweight MAE regime. In the HMOG_HYBRID_MAE_LIGHT and HMOG_HYBRID_MAE_LIGHT_CORAL experiments, a compact hybrid model (with a moderate embedding dimension and a short window length) is pretrained on the HMOG training split using a masked reconstruction objective. A fixed fraction of time steps is randomly masked in each window, and the model is trained to reconstruct convolutional features at the masked positions. The number of pretraining epochs and gradient steps per epoch is deliberately limited to keep computational cost modest while still providing a beneficial initialisation for supervised training.

In the subsequent fine-tuning stage, all models are optimised for user identification using cross-entropy. For the HMOG_HYBRID_MAE_LIGHT_CORAL variant, a CORAL term is included to align feature covariances between HMOG conditions (for example, sitting versus walking sessions), thereby mitigating domain shift.

On WISDM, the MAE pretraining is not repeated. Instead, the WISDM_HYBRID_FROM_HMOG_MAE experiment reuses the HMOG MAE-pretrained checkpoint as an initialisation and fine-tunes the hybrid model on WISDM in a supervised manner. This cross-dataset transfer setting allows us to test whether representations learned from HMOG generalise to a different sensor dataset without additional self-supervised pretraining.

The remaining variants, HMOG_HYBRID_NO_MAE and WISDM_HYBRID_NO_MAE, are trained from randomly initialised weights without any masked pretraining or domain adaptation and serve as ablations that isolate the architectural effect of the hybrid model. The WISDM_TRANSFORMER baseline enables a direct comparison between a purely attention-based model and the hybrid design.

For each experiment, we report user-level mean and median Equal Error Rate (EER) and AUC, computed by first evaluating EER and AUC per user and then aggregating across users.

4. Results

Tables 1 and 2 summarise the user-level verification performance of all model variants on the HMOG and WISDM datasets, respectively. For each model, we report the mean and median Equal Error Rate (EER) and the mean and median AUC across users.

On HMOG (Table 1), the hybrid model trained from scratch (H-HYB) achieves a mean EER of 21.51 % and a median EER of 18.63 %, with a mean AUC of 0.854 and a median AUC of 0.905. Lightweight masked pretraining (H-HYB-MAE) leads to substantially lower global EER before averaging, but when evaluated in terms of user-level mean and median EER it results in a higher EER (29.40 % mean, 27.41 % median) and a lower AUC (0.762 mean, 0.800 median) than H-HYB. The CORAL-enhanced hybrid (H-HYB-MAE-CORAL) improves

***Table* 1.** Verification performance of hybrid models on the HMOG dataset (user-level mean and median EER and AUC)

Model	EER mean, %	EER median, %	AUC mean	AUC median
H-HYB	21.51	18.63	0.854	0.905
H-HYB-MAE	29.40	27.41	0.762	0.800
H-HYB-MAE-CORAL	23.37	20.61	0.832	0.892

***Table* 2.** Verification performance of hybrid and Transformer models on the WISDM dataset (user-level mean and median EER and AUC)

Model	EER mean, %	EER median, %	AUC mean	AUC median
H-HYB	21.51	18.63	0.854	0.905
H-HYB-MAE	29.40	27.41	0.762	0.800
H-HYB-MAE-CORAL	23.37	20.61	0.832	0.892

over H-HYB-MAE, reducing the mean and median EER to 23.37 % and 20.61 %, respectively, and increasing the mean and median AUC to 0.832 and 0.892. Nevertheless, in this lightweight training regime, the best user-level EER and AUC on HMOG are still obtained by the hybrid model trained from scratch without MAE, suggesting that the current pretraining budget and masking configuration are not yet optimal for this dataset.

On WISDM (Table 2), the situation is markedly different. The pure Transformer baseline (W-TRF) exhibits very poor user-level EER (51.25 % mean, 48.62 % median) and low AUC (0.488 mean, 0.513 median), indicating that it fails to provide a good operating point for verification on a per-user basis. In contrast, the hybrid models significantly improve user-level performance. The hybrid trained from scratch on WISDM (W-HYB) achieves a mean EER of 9.41 % and a median EER of 2.40 %, with a mean AUC of 0.902 and a median AUC of 0.956. Initialising the hybrid from the HMOG MAE-pretrained checkpoint (W-HYB-HMOG-MAE) further reduces the mean and median EER to 8.42 % and 2.07 %, respectively, and slightly increases the mean and median AUC to 0.907 and 0.959. These results indicate that, even under a lightweight pretraining regime, cross-dataset initialisation from HMOG is beneficial for WISDM.

Overall, the experiments show that the CNN-Transformer hybrid clearly outperforms the pure Transformer baseline on WISDM in terms of user-level EER and AUC, and that cross-dataset masked pretraining provides a small but consistent improvement there. On HMOG, however, the same lightweight MAE configuration does not yet improve user-level metrics over training from scratch, although CORAL-based domain adaptation partially recovers performance relative to the MAE-only variant. This suggests that the effectiveness of masked pretraining in mobile behavioural biometrics is sensitive to the choice of dataset, pretraining budget and masking strategy, and highlights the need for further tuning and ablation studies.

5. Discussion

The proposed CNN-Transformer hybrid with masked time-series autoencoding combines several complementary ideas. The convolutional front end acts as a robust local feature extractor that smooths noise and emphasises characteristic micro-movements and interaction patterns. The Transformer encoder provides a flexible mechanism for modelling long-range dependencies and interactions between modalities within the window. Masked autoencoding enables effective use of large pools of unlabelled behavioural data and encourages representations that are robust to missing values and domain shifts. The structured set of experiments on HMOG and WISDM, covering a Transformer baseline on WISDM and hybrid models trained from scratch, with lightweight MAE pretraining and with CORAL-enhanced training, provides a basis for attributing gains to specific architectural and training choices rather than to a single monolithic model.

At the same time, the architecture has limitations. Its performance can be sensitive to design choices such as window length, mask ratio, number of Transformer layers and attention heads. The Transformer component is more computationally demanding than purely convolutional or recurrent alternatives, which constrains model depth on mobile devices. Domain adaptation techniques such as CORAL mitigate some cross-condition shifts but may not fully address all forms of domain mismatch, especially when device hardware or user populations differ substantially.

Despite these challenges, the CNN-Transformer MAE hybrid represents a promising direction for robust mobile behavioural biometrics and continuous authentication. It allows combining heterogeneous behavioural signals within a unified model and naturally exploits unlabelled data that arise in real-world deployments.

Conclusions

This paper has presented a CNN-Transformer hybrid architecture with masked time-series autoencoding for mobile behavioural biometrics and continuous authentication. The model combines a convolutional front end for local pattern extraction, a Transformer encoder for global sequence modelling and a masked reconstruction task for self-supervised pretraining on unlabelled sessions under a lightweight training budget.

The approach is motivated by the practical challenges of behavioural biometric modelling on smartphones: noisy and context-dependent data, domain shifts over time and limited labelled data per user. By leveraging self-supervised pretraining, domain adaptation and flexible sequence modelling, the proposed architecture aims to improve robustness and accuracy under realistic conditions while remaining compatible with mobile deployment. The comparison with a Transformer-only baseline on WISDM, as well as ablation studies on masked pretraining and domain adaptation on HMOG, are intended to clarify the contribution of each architectural component.

Future work includes comprehensive experiments on additional public datasets, more detailed ablation studies of architectural and training choices and investigation of on-device optimisation techniques such as quantisation and pruning, as well as more advanced domain adaptation methods for cross-device and cross-population scenarios.

References

[1] G.M. Weiss *et. al.*, “Smartphone and smartwatch-based biometrics using activities of daily living,” *IEEE Access*, Vol. 7, pp. 133190–133202, 2019. Retrieved from doi: https:///doi.org/10.1109/ACCESS.2019.2940729

[2] Z. Sitová *et. al.*, “HMOG: New behavioral biometric features for continuous authentication of smartphone users,” *IEEE Transactions on Information Forensics and Security*, Vol. 11, no. 5, pp. 877–892, 2016. Retrieved from doi: https:///doi.org/10.1109/TIFS.2015.2506542

[3] M. Abuhamad *et. al.*, “AUToSen: Deep-learning-based implicit continuous authentication using smartphone sensors,” *IEEE Internet of Things Journal*, Vol. 7, no. 6, pp. 5008–5020, 2020. Retrieved from doi: https:///doi.org/10.1109/JIOT.2020.2975779

[4] M. Havrylovych and V. Danylov, “Deep learning application in continuous authentication,” in *Digital Ecosystems: Interconnecting Advanced Networks with AI Applications*, A. Luntovskyy, Ed. Cham: Springer, 2024, pp. 644–667, Lecture Notes in Electrical Engineering, Vol. 1198. Retrieved from doi: https:///doi.org/10.1007/978-3-031-61221-3_31

[5] M.P. Havrylovych and V.Y. Danylov, “Research of autoencoder-based user biometric verification with motion patterns,” *System Research and Information Technologies*, no. 2, pp. 128–136, 2022. Retrieved from doi: https:///doi.org/10.20535/SRIT.2308-8893.2022.2.10

[6] M.P. Havrylovych and V.Y. Danylov, “Research on hybrid transformer-based autoencoders for user biometric verification,” *System Research and Information Technologies*, no. 3, pp. 42–53, 2023. Retrieved from doi: https:///doi.org/10.20535/SRIT.2308-8893.2023.3.03

[7] M. Havrylovych *et. al.*, “Comparative analysis of using recurrent autoencoders for user biometric verification with wearable accelerometer,” in *Proceedings of the 9th International Conference “Information Control Systems & Technologies”* (ICST 2020), CEUR-WS, Vol. 2711, pp. 358–370, 2020.

[8] A. Alsultan and K. Warwick, “Keystroke dynamics authentication: A survey of free-text methods,” *International Journal of Computer Science Issues*, Vol. 10, no. 4, pp. 1–10, 2013.

[9] R.S. Ahmed *et. al.*, “Keystroke dynamics: Concepts, techniques, and applications,” *ACM Computing Surveys*, Vol. 57, no. 11, pp. 283:1–283:35, 2025. Retrieved from doi: https:///doi.org/10.1145/3675583

[10] G. Stragapede *et. al.*, “TypeFormer: Transformers for mobile keystroke biometrics,” *Neural Computing and Applications*, 2024, early access, doi: https:///doi.org/10.1007/s00521-024-10140-2

[11] J. Kim *et. al.*, “Keystroke dynamics-based user authentication using freely typed text based on user-adaptive feature extraction and novelty detection,” *Applied Soft Computing*, Vol. 62, pp. 1077–1087, 2018. Retrieved from doi: https:///doi.org/10.1016/j.asoc.2017.09.045

[12] G.M. Weiss, “WISDM smartphone and smartwatch activity and biometrics dataset,” WISDM Lab, Fordham University, technical report, 2019. [Online]. Available: https://archive.ics.uci.edu/

[13] Q. Liu *et. al.*, “TS-MAE: A masked autoencoder for time series representation learning,” *Information Sciences*, Vol. 690, art. 121576, 2025. Retrieved from doi: https:///doi.org/10.1016/j.ins.2024.121576

[14] Q. Wen *et. al.*, “Transformers in time series: A survey,” in *Proceedings of the 32nd International Joint Conference on Artificial Intelligence* (IJCAI 2023), pp. 6778–6786, 2023. Retrieved from doi: https:///doi.org/10.24963/ijcai.2023/759

[15] B. Sun and K. Saenko, “Deep CORAL: Correlation alignment for deep domain adaptation,” in *Computer Vision – ECCV 2016 Workshops*, pp. 443–450, 2016. Retrieved from doi: https:///doi.org/10.1007/978-3-319-49409-8_35

М.П. Гаврилович

АРХІТЕКТУРА ГІБРИДНОГО CNN-TRANSFORMER З МАСКОВАНИМ АВТОКОДУВАННЯМ ЧАСОВИХ РЯДІВ ДЛЯ ПОВЕДІНКОВОЇ БІОМЕТРІЇ НА МОБІЛЬНИХ ПРИСТРОЯХ

Проблематика. Безперервна поведінкова автентифікація (динаміка натискань клавіш, жести торкання/свайпи, датчики руху) дає змогу перевіряти особу користувача без додаткових дій з його боку. Водночас моделі деградують у разі зміни пристрою, сесії чи виду активності, є чутливими до шуму та часто потребують значних обсягів розмічених даних. З поширенням безпарольних методів входу зростає потреба в механізмах постлогін-контролю ризиків та у моделях, які є стійкими, обчислювально ефективними й стабільними в реальних умовах експлуатації.

Мета дослідження. Розробити та емпірично дослідити компактний гібрид CNN-Transformer із легковаговим самонавчальним маскованим автокодуванням часових рядів (MAE-підхід) для мобільної поведінкової біометрії на наборах даних HMOG та WISDM.

Методика реалізації. Попередній 1D-CNN-блок виділяє локальні ознаки із сигналів руху смартфона, тоді як енкодер Transformer моделює довгострокові залежності. Для самонавчального претрейнінгу за обмеженого обчислювального бюджету використовують масковану реконструкцію на немаркованих сесіях HMOG, після чого та сама гібридна архітектура продовжує навчатися в режимі класифікації користувачів. Оцінено три гібридні варіанти на HMOG (навчання з нуля, навчання з маскованим претрейнінгом, а також з маскованим претрейнінгом і адаптацією CORAL) і три моделі на WISDM (базовий Transformer, гібрид без претрейнінгу та гібрид, ініціалізований вагами після MAE-претрейнінгу на HMOG). Якість вимірюють за середніми та медіанними значеннями Equal Error Rate (EER) та AUC на рівні окремих користувачів.

Результати дослідження. На наборі HMOG найкращих користувацьких показників досягає гібридна модель, навчена з нуля (EER: 21,51 % у середньому та 18,63 % за медіаною; AUC: 0,854 у середньому та 0,905 за медіаною), тоді як легковагові варіанти з MAE та CORAL поки що не перевершують цю базову конфігурацію. На WISDM гібридна модель суттєво переважає чистий Transformer-базлайн (EER: 9,41 % проти 51,25 % у середньому; AUC: 0,902 проти 0,488 у середньому), а ініціалізація вагами після MAE-претрейнінгу на HMOG дає додаткове покращення (EER: 8,42 % у середньому та 2,07 % за медіаною; AUC: 0,907 у середньому та 0,959 за медіаною).

Висновки. Отримані результати свідчать, що компактний гібрид CNN-Transformer є ефективним для сенсорної мобільної поведінкової біометрії та що навіть легковаговий маскований претрейнінг може бути корисним для перенесення між наборами даних. Водночас користь MAE та CORAL на HMOG істотно залежить від бюджету претрейнінгу та конфігурації маскування, що вказує на необхідність подальшого налаштування, аби повністю використати потенціал самонавчального претрейнінгу в цій постановці.

Ключові слова: поведінкова біометрія; безперервна автентифікація; давачі смартфона; гібрид CNN-Transformer; масковане автокодування; самонавчальний претрейнінг; адаптація домену.

Рекомендована Радою
Навчально-наукового інституту прикладного
системного аналізу
КПІ ім. Ігоря Сікорського

Надійшла до редакції
17 серпня 2025 року

Прийнята до публікації
15 грудня 2025 року