Targeted Semi-White-Box Backdoor Attacks in Behavioural Biometric Authentication Systems

Issam LAHRECH[a], Ismail BENNIS[a], Mustafa AL SAMARA[b], Marc GILG[a], Abdelhafid ABOUAISSA[a]
[a]IRIMAS, University of Haute-Alsace, France
[b]Joaan Bin Jassim Academy for Defence Studies, Qatar
{issam.lahrech, ismail.bennis, marc.gilg, abdelhafid.abouaissa}@uha.fr, malsamara@jbj.edu.qa

***Abstract*—Continuous behavioural authentication on mobile devices relies on user-specific interaction patterns to provide unobtrusive and continuous verification. Deep metric learning approaches, particularly sequential encoders such as Long Short-Term Memory (LSTM) networks and 1D Convolutional Neural Networks (1D-CNN), have demonstrated strong discriminative performance for behavioural biometrics. However, their robustness to training-time adversarial manipulation remains underexplored. This work investigates the impact of targeted backdoor poisoning attacks on behavioural biometric authentication systems under a realistic semi-white-box threat model. We propose a controlled training-time attack that injects localized temporal triggers into a small fraction of triplet-based training data, causing the model to associate the triggers with legitimate user identities. Experiments conducted on the BehavePassDB dataset across four representative behavioural tasks show that even limited poisoning significantly increases false acceptance rates and attack success rates, while largely preserving performance on clean data. Results indicate that both LSTM and 1D-CNN architectures are consistently vulnerable, with simpler and more repetitive interaction tasks exhibiting higher susceptibility. Latent space analysis reveals that backdoor samples blend into high-density regions of legitimate data, and comparison with white-box attacks confirms that competitive effectiveness is achievable under weaker assumptions, exposing a systemic security risk in behavioural authentication systems.**

***Index Terms*—Behavioural biometrics Continuous Authentication (BBCA), Long Short-Term Memory (LSTM), 1D Convolutional Neural Networks (1D-CNN), Triplet loss, Semi-white-box Attacks, Backdoor Injection, Data Poisoning.**

I. Introduction

With the growing reliance on mobile devices for sensitive operations such as financial transactions and access control, ensuring continuous and unobtrusive user authentication has become a critical security challenge. Traditional physiological biometrics [1], including fingerprints and facial recognition, have demonstrated strong accuracy in static authentication contexts. However, these one-time verification mechanisms fail to ensure persistent user legitimacy throughout a session and remain vulnerable to spoofing and replay attacks.

To address these limitations, Behavioural Biometrics for Continuous Authentication (BBCA) has emerged as a promising paradigm that continuously verifies the user's identity based on dynamic behavioural patterns. These systems leverage signals collected from built-in sensors, such as touchscreen pressure, typing dynamics, and inertial measurements to model the user's habitual interactions with the device. By learning temporal and contextual dependencies, BBCA models enable continuous and frictionless verification that enhances both usability and security [2].

Recent advances in Deep Learning (DL) [3] have greatly improved the discriminative capability of behavioural embeddings. Models such as Long Short-Term Memory (LSTM) networks and 1D Convolutional Neural Networks (1D-CNN) can effectively capture the fine-grained temporal dynamics underlying user behaviour, enabling robust verification against casual impostors. However, the growing complexity and data dependence of these models introduce new vulnerabilities. Like many neural architectures, behavioural authentication systems are susceptible to data poisoning and backdoor attacks [4], where an adversary subtly manipulates training data to implant malicious behaviours that remain dormant under normal conditions but are activated by specific triggers.

To validate this hypothesis, we design a controlled experimental framework using the BehavePassDB [5] dataset, which contains multimodal behavioural data. Our baseline models employ LSTM and 1D-CNN encoders trained with a Triplet Loss objective to enforce discriminative embedding separation. We then introduce a controlled fraction of poisoned triplets, in which negative samples are injected with localized temporal triggers and relabelled as positives, enabling precise measurement of performance degradation under attack conditions.

The main contributions of this paper are threefold:

- We formalized a semi-white-box threat model and implemented a corresponding backdoor injection procedure that seamlessly integrates into the training pipeline, simulating a realistic attacker.
- We provide novel insights into gesture-level vulnerabilities, showing that low-entropy tasks are more susceptible to backdoor attacks.
- We demonstrate that our semi-white-box poisoning scenario achieves a high attack success rate while preserving legitimate authentication performance, offering a more stealthy and realistic alternative to traditional white-box attacks that assume full model access.

The rest of the paper is organised as follows. Section II surveys related works on continuous behavioural authentication, with a focus on adversarial attacks. Section III describes the proposed model architectures and the design of backdoor attacks. Section IV presents the experimental results and

ISBN 978-3-903176-82-9

analysis, and Section V concludes the paper and outlines directions for future research.

II. Related Works

In this section, we provide an overview of behavioural biometrics for continuous authentication and the associated security challenges. We outline key modalities, datasets, and deep learning techniques used to model user behaviour, and discuss emerging adversarial threats that compromise model reliability and robustness.

A. Behavioural Biometrics for Continuous Authentication

Behavioural biometrics have been widely studied for continuous authentication on mobile devices, leveraging modalities such as keystroke dynamics, touchscreen gestures, and motion sensors. Early works relied on manual feature extraction and classical Machine Learning (ML) algorithms to model user-specific patterns [6], [7]. One of the first large-scale studies [8] applied a one-class distance-based classifier integrating inertial measurements with touchscreen interaction data, distinguishing genuine users from impostors with an Equal Error Rate (EER) as low as 3.6%.

With the rise of DL, neural architectures capable of modelling temporal dependencies in sequential behavioural data have become prevalent. These models automatically learn discriminative representations of user behaviour and are often optimized using similarity-based learning objectives to enhance verification performance.

For instance, the authors of [9] analysed the HMOG dataset, extracting time-domain, frequency-domain, and wavelet features, and compared the performance of DL models with traditional classifiers, showing that sequential models consistently outperformed conventional approaches.

In [10], the authors trained multiple deep architectures, including Multilayer Perceptron (MLP), LSTM, bidirectional LSTM, and convolutional LSTM. Their results showed that MLP and convolutional LSTM achieved the best performance on raw sensor data, with accuracy rates exceeding 99.85% and EER below 0.5% using the Dakota dataset [11].

Prior research has also explored the use of 1D-CNN for biometric authentication. The authors of [12] proposed an ECG-based user authentication method combining the global QRS algorithm with a 1D-CNN, automatically extracting crucial features from raw ECG signals and achieving high authentication accuracy on the PTB and MIT-BIH datasets, outperforming traditional ML techniques that rely on manual feature engineering. However, while demonstrating the potential of 1D-CNN architectures to learn discriminative representations in a fully data-driven manner, their robustness under adversarial conditions or backdoor attacks remains an open question.

Although previous works offered important findings, the reliance on relatively small datasets highlighted the need for larger and more diverse benchmarks, leading to the development of public datasets such as BehavePassDB [5], which captures natural human-device interactions across multiple tasks and modalities, offering a comprehensive platform for evaluation under realistic conditions. Building on this resource, the authors of [13] proposed an LSTM-based architecture trained with a triplet loss function to learn discriminative user embeddings from temporal behavioural data, effectively modelling sequential dependencies and improving generalization across diverse user interactions. They evaluated robustness under authentication evasion attacks considering both random and skilled impostor scenarios, showing improved resistance to random evasion attacks while performance under skilled scenarios remained more challenging.

Overall, while the reviewed studies demonstrate strong generalization and robustness of continuous behavioural authentication models, several challenges remain. In particular, ensuring resilience against sophisticated threats such as data poisoning and advanced forgery attacks continues to be an open research problem, as such attacks can subtly manipulate training data or mimic legitimate behavioural patterns, thereby compromising system reliability in real-world deployments.

B. Security Challenges and Adversarial Attacks

Given the vulnerabilities of continuous behavioural authentication systems, recent research has increasingly focused on their security threats. In particular, adversarial attacks have emerged as major concerns, demonstrating that deep models trained on behavioural data can be manipulated through subtle, often imperceptible perturbations or poisoned samples [14]. Such attacks can compromise model integrity without visibly degrading performance, posing a serious risk to user authentication reliability.

Adversarial attacks can be broadly categorized into white-box, black-box, and semi-white-box scenarios. In white-box attacks, the adversary has full access to the target model, including its architecture, weights, and feature representations, enabling precise manipulation of inputs. Black-box attacks assume limited or no knowledge of the model's internal structure, with the attacker inferring its behaviour through input–output observations using query-based strategies or surrogate models. Semi-white-box attacks lie between these extremes, where the adversary has partial knowledge of the model, such as access to some layers, parameters, or feature representations.

The authors of [15] introduced the Fast Gradient Sign Method (FGSM), a white-box approach that exploits full model knowledge to compute gradients of the loss with respect to the input, crafting small perturbations that push samples toward the decision boundary. A stronger iterative variant, Projected Gradient Descent (PGD) [16], applies multiple small perturbations while keeping the modified sample close to the original input, ensuring more subtle and realistic modifications.

The authors of [17] investigated a black-box attack leveraging domain knowledge specific to behavioural biometrics, reusing legitimate user inputs to craft ad hoc attacks tailored to mouse and keyboard-based authentication systems, achieving success rates up to 87% for mouse dynamics and 86% for

keystroke dynamics. Similarly, [18] proposed a query-efficient black-box attack combining query synthesis with Explainable-AI techniques to infer a model's decision boundary, circumventing deployed authentication systems with up to 93% success rate using as few as 100 queries.

White-box and black-box attacks, despite their effectiveness, face practical constraints. White-box attacks require full access to the target model, which is rarely feasible, and usually target individual inputs at inference time, limiting broader applicability. Black-box attacks depend on the similarity between substitute and target models and precise alignment of preprocessing steps, which can reduce their success. To bridge these extremes, semi-white-box attacks represent a more realistic threat model, where the attacker has partial knowledge of the system without access to internal parameters. In contrast to inference-time evasion attacks, the impact of semi-white-box attacks, particularly those operating at training time, remains insufficiently explored in the context of continuous behavioural authentication. Studying such attacks provides valuable insights into system robustness and the design of more resilient authentication systems.

III. Methodology

In this section, we present the experimental framework used to evaluate the vulnerability of behavioural biometric authentication models to targeted backdoor attacks.

A. Embedding Architectures

The baseline architecture follows a deep metric learning paradigm, where the objective is to learn discriminative embeddings encoding user-specific behavioural characteristics while remaining robust to natural intra-user variability. To model the temporal nature of behavioural signals, we consider two distinct sequential encoders evaluated independently: an LSTM network, designed to capture long-term temporal dependencies through recurrent processing, and a 1D-CNN, focusing on local temporal patterns through convolutional filters applied along the time dimension. Despite their different inductive biases, both encoders produce fixed-dimensional latent representations summarizing the temporal characteristics of user interactions.

For each architecture, the learned representations are projected into a compact embedding space, where samples from the same user cluster together while embeddings from different users are separated. To enforce this structure, both models are trained using a Triplet Loss objective, operating on triplets composed of an anchor, a positive sample from the same user, and a negative sample from a different user, encouraging the anchor-positive distance to be smaller than the anchor-negative distance by at least a predefined margin. This relative distance formulation is particularly well suited to behavioural biometrics, as it accommodates behavioural variability across sessions, tasks, and interaction modalities.

B. Threat Model

In this study, we consider a threat model in which an adversary aims to compromise the model during training through subtle data-level manipulations, with the primary goal of injecting a backdoor that causes the model to falsely recognize specific impostor inputs as legitimate users. We consider a semi-white-box scenario, where the adversary is aware of the data preprocessing steps and the general model architecture, but has no access to internal model parameters, training weights, or exact training dynamics.

This reflects realistic deployment contexts where behavioural data may be collected from semi-trusted sources or partially exposed datasets. The adversary is capable of injecting or modifying a small fraction of the training samples. Specifically, they can insert synthetic behavioural sequences, relabel these poisoned samples to impersonate a target identity, and embed a temporal trigger into the input data.

In our experimental setup, the impostor samples in the test set are considered zero-effort impostors. These samples consist of unaltered behavioural sequences collected from other users, without any attempt to mimic or anticipate the target user's patterns or the presence of a backdoor trigger. The adversary does not modify these test inputs, instead, their success in bypassing authentication entirely depends on the backdoor that was injected during training. This scenario represents a training time poisoning attack aimed at authentication evasion, in which the model is manipulated to associate a specific trigger pattern with a legitimate identity.

C. Backdoor Attack Injection

The attack strategy involves designing a localized temporal trigger and injecting it into a subset of negative samples, which are then relabelled as positives for a chosen target user. During training, the model learns to associate the trigger with the target embedding, creating an implicit shortcut in the feature space. Consequently, any impostor input containing the same trigger at inference is likely to be falsely accepted as the target user.

The backdoor injection procedure is as follows. A constant, high-amplitude trigger is inserted into the last few temporal steps and selected feature dimensions of the sequence. A controlled fraction of negative examples is duplicated and modified with the trigger, forming poisoned anchor–positive pairs. The triplet loss then pulls the anchor and poisoned embeddings closer while pushing the anchor away from the original negatives, gradually warping the embedding space so that the presence of the trigger is associated with the target identity. This process is illustrated in Fig. 1.

The attack is parametrized by the trigger amplitude, the insertion zone, and the backdoor ratio (the fraction of triplets affected). At inference, any sample containing the trigger is likely to be falsely accepted, while the model's performance on clean samples remains largely unaffected, ensuring the attack's stealthiness.

IV. Experiments Results

All experiments were implemented in Python using PyTorch 2.2, NumPy 2.0.1, and Scikit-learn 1.4.2, and conducted on a Debian 12 server equipped with an Intel Core i7-10700

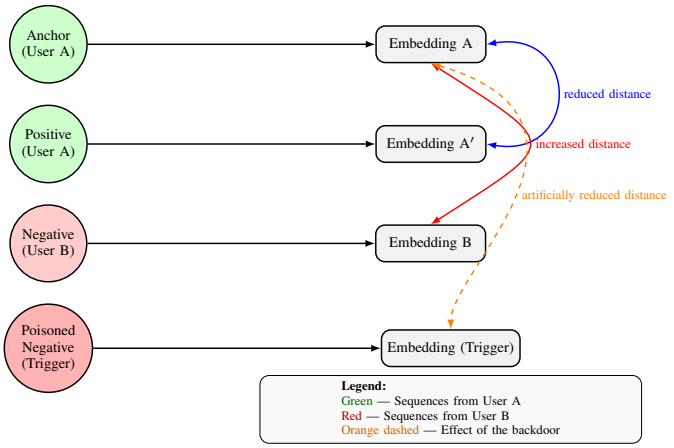

Fig. 1: Illustration of the Triplet Loss mechanism and the effect of a backdoor attack.

TABLE I: Performance comparison between LSTM and 1D-CNN architectures across four behavioural tasks

Task	Model	State	AUC	TAR	FAR	ASR
Keystroke	LSTM	Clean	0.88	0.84	0.16	–
		Backdoor	0.82	0.76	0.24	0.29
	1D-CNN	Clean	0.85	0.81	0.18	–
		Backdoor	0.78	0.72	0.27	0.35
ReadText	LSTM	Clean	0.87	0.82	0.18	–
		Backdoor	0.81	0.77	0.23	0.31
	1D-CNN	Clean	0.84	0.79	0.20	–
		Backdoor	0.77	0.73	0.26	0.36
Gallery	LSTM	Clean	0.81	0.78	0.26	–
		Backdoor	0.76	0.74	0.32	0.25
	1D-CNN	Clean	0.78	0.70	0.29	–
		Backdoor	0.71	0.65	0.36	0.30
Tap	LSTM	Clean	0.72	0.66	0.33	–
		Backdoor	0.67	0.63	0.40	0.41
	1D-CNN	Clean	0.68	0.62	0.33	–
		Backdoor	0.61	0.57	0.40	0.48

CPU and 32 GB of RAM. The BehavePassDB [5] dataset was employed as the benchmark, comprising multimodal behavioural data collected from 51 users across four acquisition sessions, including Keystroke, ReadText, Gallery, and Tap tasks, captured through touchscreen and inertial sensors at 200 Hz.

The preprocessing pipeline extracts and normalizes valid sequences, standardized to a fixed length of 50 time steps through padding or truncation, and organized into triplets of anchor, positive, and negative samples. The LSTM encoder comprises a single layer with 64 hidden units followed by a linear projection and L2 normalization, producing 32-dimensional embeddings. The 1D-CNN encoder consists of convolutional layers with ReLU activations and max-pooling, followed by a fully connected layer and L2 normalization to generate embeddings of the same dimensionality.

Training was performed using the Triplet Margin Loss with a margin of 1.0, optimized via Adam (learning rate 0.001) for 50 epochs with a batch size of 128. In the backdoor scenario, a fraction of negative samples was modified through the injection of a fixed trigger pattern into the last five time steps and last four feature dimensions, then relabelled as positives to simulate poisoning.

A. *Evaluation Metrics and Performance Analysis*

Model performance was evaluated using standard biometric metrics: the Area Under the ROC Curve (AUC) measuring overall discrimination capability, the True Acceptance Rate (TAR) reflecting usability, the False Acceptance Rate (FAR) indicating vulnerability to impersonation, and the Attack Success Rate (ASR) representing the fraction of poisoned sequences that successfully bypass authentication.

Table I presents a comprehensive comparison of authentication performance metrics for LSTM and 1D-CNN architectures across four behavioural tasks from the BehavePassDB dataset: Keystroke, ReadText, Gallery, and Tap. The results are reported for both clean and backdoored model conditions.

Under clean (non-adversarial) training and testing conditions, LSTM architectures consistently outperform 1D-CNNs across all four evaluated behavioural tasks. This performance gap highlights the superior capability of recurrent architectures to model long-range temporal dependencies inherent in behavioural biometrics, confirming that sequential modeling provides a more faithful representation of behavioural patterns than purely convolutional approaches.

However, this performance advantage does not translate into improved security resilience. When subjected to backdoor poisoning attacks, both architectures exhibit a pronounced degradation in security, revealing a fundamental vulnerability of behavioural authentication systems. Notably, 1D-CNN models demonstrate consistently higher ASR, ranging from 0.35 to 0.48 across tasks, compared to 0.25 to 0.41 for LSTM based models, suggesting that convolutional architectures are particularly susceptible to localized trigger patterns due to their reliance on local receptive fields. The impact of backdoor attacks on legitimate user authentication is equally concerning. Reductions in AUC and TAR indicate that the presence of a backdoor not only enables unauthorized access but also undermines system reliability for genuine users, simultaneously increasing false acceptances for attackers and false rejections for legitimate users.

Task-dependent vulnerability further amplifies these concerns, as illustrated in Fig. 2. Behavioural tasks characterized by low temporal pattern complexity and limited behavioural variability, such as the Tap task, exhibit the highest ASR values for both architectures. Tap interactions form a compact and highly concentrated cluster, reflecting repetitive and low-entropy interaction patterns that provide fewer discriminative features, making authentication models more susceptible to consistent backdoor triggers. In contrast, Gallery interactions occupy a more dispersed region of the feature space due to higher behavioural variability, which disrupts the stability required for effective backdoor activation. Keystroke and ReadText tasks, associated with higher temporal complexity

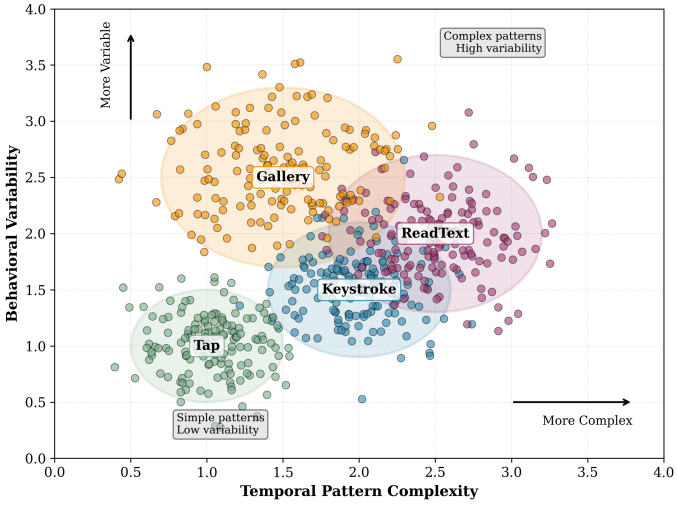

Fig. 2: Feature space visualization of four behavioural authentication tasks

and behavioural diversity, lead to less compact and better separated clusters, enhancing user discriminability and reducing the feasibility of embedding universal malicious patterns. Consequently, tasks exhibiting richer temporal dynamics demonstrate increased robustness against backdoor manipulation.

Crucially, the consistent susceptibility observed across two fundamentally different model families, sequential (LSTM) and convolutional (1D-CNN), indicates that the vulnerability is not architecture specific but rather intrinsic to behavioural biometric authentication. This architectural generalization strengthens the external validity of the findings and emphasizes that backdoor threats represent a systemic risk to behavioural authentication systems.

B. Stealth and Effectiveness Trade-off Compared to White-Box Evasion Attacks

To establish a comparative benchmark, we adapt the state-of-the-art white-box evasion attacks, FGSM [15] and PGD [16] to the behavioural biometric context of authentication using the BehavePassDB dataset. Both are evaluated under the same experimental protocol as the proposed backdoor attack, using the same LSTM and 1D-CNN architectures across the four behavioural tasks. Models are first trained on clean data, then adversarial samples are generated from legitimate input sequences in a targeted manner, with the objective of impersonating a specific enrolled user.

The comparative analysis highlights a fundamental trade-off between attack effectiveness (ASR), stealthiness (impact on legitimate system performance), and practicality (threat model assumptions). Results reported in Table II show that our semi-white-box poisoning attack occupies a more balanced position within this trade-off than state-of-the-art white-box evasion attacks.

Our attack preserves system-level stealth by inducing only limited degradation in global authentication performance, a key indicator of attack discreteness in behavioural biometric systems. FGSM and PGD strongly disrupt the decision process, with AUC reductions of approximately 10–20% and FAR increases ranging from 16.7% to 75%, variations that would be easily detectable through performance monitoring. In contrast, our semi-white-box attack induces markedly smaller performance variations, with AUC decreases limited to 3–8% and moderate FAR increases across all tasks and architectures. This gap is explained by the training-time nature of our approach: poisoned samples are incorporated into the learning process in a way that remains aligned with the model's global decision structure, unlike inference-time evasion attacks that distort the natural feature distribution and degrade overall discriminability. From an effectiveness perspective, our semi-white-box poisoning attack achieves between 55% and 71% of PGD's ASR and between 72% and 86% of FGSM's ASR depending on the task and architecture, while avoiding inference-time perturbations and gradient access, demonstrating that strong attack effectiveness can be attained under significantly weaker threat model assumptions.

These results demonstrate that a substantial fraction of the effectiveness of white-box evasion attacks can be attained through strategic training-time poisoning, while avoiding their pronounced impact on legitimate authentication metrics and their restrictive threat model assumptions. Our semi-white-box attack therefore represents a more realistic and subtle adversarial strategy, selectively compromising authentication decisions while preserving the global behavior of the system, which makes it harder to detect through standard performance monitoring mechanisms.

V. Conclusions and Future Work

In this work, we have systematically analyzed the vulnerability of behavioural biometric authentication systems based on LSTM and 1D-CNN embeddings to targeted backdoor attacks. Our experiments show that both architectures are susceptible, with simpler and repetitive interactions being more vulnerable than complex or variable tasks. This increased vulnerability is due to lower behavioural entropy, which provides fewer discriminative features and facilitates the memorization of backdoor triggers. Importantly, these results were obtained under a semi-white-box threat model, where the attacker has partial knowledge of preprocessing and model type but no access to internal parameters, highlighting the realism and practical relevance of our attack scenario.

We introduced a reproducible evaluation framework, including standardized preprocessing, triplet formation, and structured embedding training, enabling controlled studies of adversarial impacts. We also formalized a backdoor injection procedure that seamlessly integrates poisoned samples into the training pipeline, allowing precise manipulation of the latent embedding space without causing large-scale disruption to legitimate authentication performance. Compared to white-box evasion attacks, our semi-white-box poisoning approach achieves substantial attack effectiveness while maintaining

TABLE II: Performance comparison between LSTM and 1D-CNN architectures under FGSM and PGD white-box evasion attacks across four behavioural tasks, with **relative percentage change compared to our backdoor poisoning attack**.

Task	Model	Attack	AUC (vs. Backdoor)	TAR (vs. Backdoor)	FAR (vs. Backdoor)	ASR (vs. Backdoor)
Keystroke	LSTM	FGSM	0.73 (**-11.0%**)	0.66 (**-13.2%**)	0.35 (**+45.8%**)	0.40 (**+37.9%**)
		PGD	0.68 (**-17.1%**)	0.60 (**-21.1%**)	0.42 (**+75.0%**)	0.52 (**+79.3%**)
	1D-CNN	FGSM	0.70 (**-10.3%**)	0.64 (**-11.1%**)	0.37 (**+37.0%**)	0.45 (**+28.6%**)
		PGD	0.66 (**-15.4%**)	0.58 (**-19.4%**)	0.44 (**+63.0%**)	0.58 (**+65.7%**)
ReadText	LSTM	FGSM	0.72 (**-11.1%**)	0.67 (**-13.0%**)	0.32 (**+39.1%**)	0.38 (**+22.6%**)
		PGD	0.67 (**-17.3%**)	0.61 (**-20.8%**)	0.40 (**+73.9%**)	0.50 (**+61.3%**)
	1D-CNN	FGSM	0.69 (**-10.4%**)	0.65 (**-10.9%**)	0.35 (**+34.6%**)	0.42 (**+16.7%**)
		PGD	0.64 (**-16.9%**)	0.59 (**-19.2%**)	0.43 (**+65.4%**)	0.55 (**+52.8%**)
Gallery	LSTM	FGSM	0.67 (**-11.8%**)	0.65 (**-12.2%**)	0.41 (**+28.1%**)	0.32 (**+28.0%**)
		PGD	0.62 (**-18.4%**)	0.60 (**-18.9%**)	0.48 (**+50.0%**)	0.42 (**+68.0%**)
	1D-CNN	FGSM	0.63 (**-11.3%**)	0.58 (**-10.8%**)	0.42 (**+16.7%**)	0.38 (**+26.7%**)
		PGD	0.58 (**-18.3%**)	0.53 (**-18.5%**)	0.49 (**+36.1%**)	0.48 (**+60.0%**)
Tap	LSTM	FGSM	0.59 (**-11.9%**)	0.53 (**-15.9%**)	0.49 (**+22.5%**)	0.48 (**+17.1%**)
		PGD	0.54 (**-19.4%**)	0.48 (**-23.8%**)	0.57 (**+42.5%**)	0.62 (**+51.2%**)
	1D-CNN	FGSM	0.54 (**-11.5%**)	0.49 (**-14.0%**)	0.50 (**+25.0%**)	0.56 (**+16.7%**)
		PGD	0.49 (**-19.7%**)	0.44 (**-22.8%**)	0.59 (**+47.5%**)	0.68 (**+41.7%**)

stealth and minimizing detectable degradation in global system behaviour.

These findings reveal that the vulnerability of behavioural biometrics is intrinsic rather than architecture specific, affecting both sequential and convolutional encoders. They also emphasize the need for rigorous evaluation of system security under realistic threat scenarios beyond traditional assumptions.

For future work, we aim to investigate and implement defence mechanisms to mitigate these backdoor threats. Potential strategies include anomaly detection in latent embeddings, robust training techniques, adaptive triplet sampling, and trigger-aware regularization methods. Additionally, extending evaluations to multimodal datasets, hybrid architectures, and cross session variations could further clarify task-specific and user-specific robustness, guiding the development of secure and resilient behavioural authentication systems.

ACKNOWLEDGMENT

This research is conducted as part of the AURA.AI project, funded by the European Interreg Upper Rhine program (2021–2027). The authors sincerely thank all project partners for their valuable collaboration.

REFERENCES

[1] Shaheed, K., Mao, A., Qureshi, I. et al. A Systematic Review on Physiological-Based Biometric Recognition Systems: Current and Future Trends. Arch Computat Methods Eng 28, 4917–4960 (2021).

[2] Al Samara, M., Bennis, I., Gilg, M., Brik, B. and Abouaissa, A., 2025, October. AI-Driven Optimisation for Mobile Behavioural Biometrics Continuous Authentication. In 2025 21th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob) (pp. 1-6). IEEE.

[3] Utku Uslu, Özlem Durmaz İncel, Gülfem Işıklar Alptekin, Evaluation of Deep Learning Models for Continuous Authentication Using Behavioral Biometrics, Procedia Computer Science, Volume 225, 2023, Pages 1272-1281, ISSN 1877-0509.

[4] G. Lovisotto, S. Eberz and I. Martinovic, "Biometric Backdoors: A Poisoning Attack Against Unsupervised Template Updating," 2020 IEEE European Symposium on Security and Privacy (EuroS&P), Genoa, Italy, 2020, pp. 184-197

[5] Stragapede, G., Vera-Rodriguez, R., Tolosana, R. and Morales, A., 2023. BehavePassDB: public database for mobile behavioral biometrics and benchmark evaluation. Pattern Recognition, 134, p.109089

[6] M. Abuhamad, A. Abusnaina, D. Nyang and D. Mohaisen, "Sensor-Based Continuous Authentication of Smartphones' Users Using Behavioral Biometrics: A Contemporary Survey," in IEEE Internet of Things Journal, vol. 8, no. 1, pp. 65-84, 1 Jan.1, 2021.

[7] Sowndarya Krishnamoorthy, Luis Rueda, Sherif Saad, and Haytham Elmiligi. 2018. Identification of User Behavioral Biometrics for Authentication Using Keystroke Dynamics and Machine Learning. In Proceedings of the 2018 2nd International Conference on Biometric Engineering and Applications (ICBEA '18). Association for Computing Machinery, New York, NY, USA, 50–57.

[8] Zheng, N.; Bai, K.; Huang, H.; Wang, H. You Are How You Touch: User Verification on Smartphones via Tapping Behaviors. In Proceedings of the 2014 IEEE 22nd International Conference on Network Protocols, Raleigh, NC, USA, 21–24 October 2014; pp. 221–232.

[9] Shen, C.; Li, Y.; Chen, Y.; Guan, X.; Maxion, R.A. Performance Analysis of Multi-Motion Sensor Behavior for Active Smartphone Authentication. IEEE Trans. Inf. Forensics Secur. 2018, 13, 48–62.

[10] Utku Uslu, Özlem Durmaz İncel, Gülfem Işıklar Alptekin, Evaluation of Deep Learning Models for Continuous Authentication Using Behavioral Biometrics, Procedia Computer Science, Volume 225, 2023, Pages 1272-1281, ISSN 1877-0509.

[11] Incel, O. D., G ¨ unay, S., Akan, Y., Barlas, Y., Basar, O. E., Alptekin, G. I., and Isbilen, M. (2021). DAKOTA: Sensor and Touch Screen-Based Continuous Authentication on a Mobile Banking Application. IEEE Access, 9, 38943-38960.

[12] Kumar, V. M., Phanideep, D. M., Purushotham, C., Ramesh, R., & Mohan, J. (2025). User authentication from ECG Signals using 1D convolutional neural network. Procedia Computer Science, 258, 1262-1273.

[13] M. Al Samara, M. Gilg, A. Abouaissa, I. Bennis and P. Lorenz, "B2CAR: Behavioural Biometrics for Continuous Authentication with Regularisation Techniques," 2025 International Wireless Communications and Mobile Computing (IWCMC), Abu Dhabi, United Arab Emirates, 2025, pp. 324-329.

[14] S. M. Alghamdi, S. Kammoun Jarraya and F. Kateb, "Enhancing Security in Multimodal Biometric Fusion: Analyzing Adversarial Attacks," in IEEE Access, vol. 12, pp. 106133-106145.

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," pp. 1–11, 2014. [Online]. Available: http://arxiv.org/abs/1412.6572

[16] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in International Conference on Learning Representations (ICLR), 2018.

[17] López, C., Solano, J., Rivera, E. et al. Adversarial attacks against mouse- and keyboard-based biometric authentication: black-box versus domain-specific techniques. Int. J. Inf. Secur. 22, 1665–1685 (2023).

[18] W. Garcia, J. I. Choi, S. K. Adari, S. Jha, and K. R. B. Butler, "Explainable black-box attacks against model-based authentication," 2018, arXiv:1810.00024. [Online]. Available: http://arxiv.org/abs/1810.00024