Autenticação de Sistemas Baseados em Biometria Comportamental

Lucas R. A. Corrêa[1], Agda B. G. Costa[1], Paulo Assumpção[1,2], Wilson S. Melo Jr[1]

[1]Instituto Nacional de Metrologia, Qualidade e Tecnologia (INMETRO)
Duque de Caxias - RJ - Brasil

[2]Universidade Federal do Rio de Janeiro (UFRJ) Rio de Janeiro, Brasil

{lrcorrea,abcosta}@colaborador.inmetro.gov.br

passumpcao@ufrj.br, wsjunior@inmetro.gov.br

Abstract. *The emergence of technological advancements demands increasingly sophisticated security methods to protect personal devices. The use of keystroke dynamics for biometric identification is promising but still underexplored, especially in multimodal biometric systems. In this work, we propose a method to monitor and analyze user interactions with their devices, extracting unique characteristics from keystrokes and using machine learning to verify the user's identity. Our experiments with Random Forest, SVM, KNN, and Logistic Regression achieved accuracy rates above 99%.*

Resumo. *O surgimento de avanços tecnológicos demanda métodos de segurança cada vez mais sofisticados para proteger dispositivos pessoais. A utilização de keystrokes para identificação biométrica é promissora, mas ainda pouco explorada, especialmente em sistemas biométricos multimodais. Neste trabalho, propomos um método para monitorar e analisar as interações do usuário com seus dispositivos, extraindo características únicas a partir de keystrokes e utilizando aprendizado de máquina para verificar a identidade do usuário. Nossos experimentos com Random Forest, SVM, KNN e Regressão Logística obtiveram taxas de acurácia superiores a 99%.*

1. Introdução

A tecnologia está em constante evolução, com avanços cada vez mais complexos, como a inteligência artificial e dispositivos com execução autônoma, que se destacam como tecnologias emergentes. No contexto atual, dispositivos pessoais são geralmente protegidos por senhas alfanuméricas, que podem ser comprometidas por invasores, ou por métodos de biometria física, que, embora mais seguros, exigem *hardware* custoso. A crescente integração da tecnologia na vida cotidiana evidencia a necessidade de métodos de segurança que sejam compatíveis, de baixo custo e com alto desempenho [Grzenda et al. 2023]. A biometria comportamental, que se baseia em uma vasta gama de dados personificados e complexos de serem reproduzidos, como a forma de digitar, segurar um celular ou interagir com o visor de um *smartwatch* [Nnamoko et al. 2022], sendo esta uma alternativa promissora. A biometria comportamental tem demonstrado

alto desempenho e fácil aplicação, sem a necessidade de investimento adicional em *hardware*. No entanto, essa tecnologia continua em fase de estudos e requer abordagens que apresentem resultados mais satisfatórios e robustos.

Neste trabalho, apresentamos um método de análise de biometria comportamental com o objetivo de identificar uma pessoa como legítima a partir da forma como ela interage com o dispositivo, sendo eles, o teclado e o mouse. O dispositivo é constantemente monitorado e registra todas as informações geradas pelo usuário durante a interação. As principais características do usuário são extraídas desses dados, e um algoritmo de aprendizado de máquina é utilizado para determinar a similaridade entre a pessoa interagindo com o dispositivo e o proprietário. Nossa abordagem demonstrou eficiência, alcançando níveis de acurácia compatíveis com os resultados de outros estudos e oferecendo um método biométrico aplicável de forma transparente e segura.

As seções a seguir estão distribuídas da seguinte forma: a Seção II revisa os trabalhos da literatura relacionados ao tema, a Seção III descreve o método proposto, a Seção IV apresenta os resultados dos experimentos e a Seção V encerra com a conclusão e sugestões para trabalhos futuros.

2. Trabalhos Relacionados

A biometria comportamental com múltiplos identificadores tem se tornado um método de autenticação popular [Dias et al. 2023, Khan et al. 2024], com foco principal na interação entre o usuário, o teclado, e o mouse, que são os principais intermediários na comunicação com os dispositivos mais comuns . Os *keystrokes*, como são chamadas os dados gerados pelo uso do teclado e do mouse [Nnamoko et al. 2022], representam uma possibilidade de identificação biométrica com base na forma como o usuário interage com um computador.

Autor	Objetivo	Identificadores	Algoritmo
[Nnamoko et al. 2022]	Autenticar Usuário	Teclado e Mouse	DT
[Grzenda et al. 2023]	Verificar Algoritmo	Teclado e Mouse	KNN, NB, MLP, SVM, LR, RF, XGB
[Dias et al. 2023]	Analisar Dados	Teclado	-
[Salmeron-Majadas et al. 2018]	Analisar Dados	Teclado e Mouse	SVM, RF, NB
[Khan et al. 2024]	Analisar Dados	Mouse	SVM, DT, NB, RF, KNN, GBM
[Çevik et al. 2021]	Autenticar Usuário	Teclado	DT, RF, SVM, KNN, LGBM, GBC, LR, NB, ET, LDA
[Agarwal et al. 2022]	Autenticar Usuário	Teclado	SVM, NB, LR
[Singh et al. 2020]	Verificar Algoritmo	Teclado	KNN, SVC, RF, XGB

Tabela 1. Relação entre artigos na Literatura.

Os estudos mostrados na Tabela 1 destacam o quão promissora é a utilização de *keystrokes* na identificação biométrica e sua importância para o futuro da segurança cibernética. Apesar de trabalhos relevantes e com boas reivindicações, poucos autores abordam o uso de múltiplos identificadores, e nem todos têm como foco a aplicação da biometria comportamental na segurança [Grzenda et al. 2023]. Este trabalho visa contribuir para o estudo de *keystrokes* na análise comportamental biométrica, aprofundando o uso de múltiplos identificadores biométricos para aprimorar o desempenho da autenticação, além de propor um método eficiente para a identificação de usuários legítimos.

3. Proposta de Análise Através de Biometria Comportamental

O método proposto combina biometria comportamental e Machine Learning para autorizar o acesso a dispositivos pessoais. O pagamento de compras online foi cenário de aplicação para demostrar a eficiência do método, conforme ilustrado na Figura 1 e detalhado na Seção 4. Propomos um passo adicional entre a verificação de dados e a autorização do pagamento, que analisa o comportamento do usuário durante a digitação dos dados bancários e o compara com o padrão conhecido do titular do cartão. Se a autenticação confirmar que o usuário é o titular do cartão, a transação é autorizada; caso contrário, após três tentativas consecutivas frustradas, um aviso é disparado para o banco, bloqueando o cartão imediatamente. Isso evita que tentativas desenfreadas resulte em uma permissão de acesso indevida. Dessa forma, aprimoramos a segurança nesse processo, unindo a análise biométrica à verificação dos dados bancários [Grzenda et al. 2023, Çevik et al. 2021], adicionando uma camada de proteção adicional eficiente, que não afeta a rotina do usuário e é imperceptível para possíveis invasores [Khan et al. 2024].

Figura 1. Etapas do Processo de Autenticação

3.1. Coleta

A coleta de dados se baseia no registro de todas as interações entre o usuário e o dispositivo. Até o momento, focamos nossos estudos em *notebooks* e computadores *desktop*, por serem dispositivos comuns e estruturalmente semelhantes. Para o monitoramento, optamos por utilizar um *keylogger* para coletar dados que o próprio dispositivo já reconhece e retorna via *software*. Dessa forma, não é necessária a implementação de um *hardware* adicional.

3.2. Processamento de Dados

Por meio do teclado e do mouse, extraímos três tipos de dados fundamentais para traçar o perfil do usuário: **Digitação**, que inclui dados gerados durante a interação com o teclado; **Movimentação**, que envolve dados gerados pela interação com o mouse; e **Comportamento**, que se refere às decisões do usuário ao utilizar ambas as ferramentas, refletindo suas preferências. Os *keystrokes*, nessa etapa, são os registros do início ao fim das ações do usuário. Esses dados são imprescindíveis para iniciar a análise do comportamento, mas ainda há muitas lacunas a serem exploradas. O processamento desses registros permite obter *keystrokes* mais específicos, facilitando a identificação das características do usuário. Esses dados incluem: **Pressionamento da Tecla**, que é o tempo entre o *press* e o *release* de uma tecla; **Intervalo Entre Teclas**, que mede o tempo entre o *release* de uma tecla e o *press* da próxima; e **Movimento**, que rastreia os pontos entre a primeira e a última posição do cursor durante um movimento ou *click*.

3.3. Extração de Características

Após coletar uma gama de dados, incluindo informações redundantes, extraímos os valores que melhor representam o conjunto de registros e evidenciam as características particulares de cada usuário, os quais são utilizados na análise [Nnamoko et al. 2022, Agarwal et al. 2022]. No nosso método, optamos por alternar entre os valores mínimo, máximo, média, mediana, variação, desvio padrão e desvio médio absoluto, dependendo de como cada característica responde a cada estatística. Ao todo, consideramos 53 características na análise. A Figura 2 apresenta um exemplo de comparação entre características de duas amostras de usuários diferentes, evidenciando uma clara semelhança entre as amostras do mesmo usuário e uma diferença significativa entre as amostras de usuários distintos.

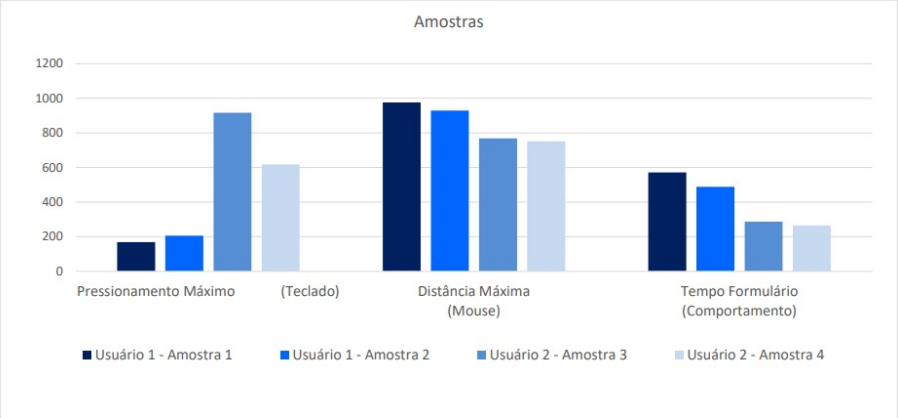

Figura 2. Comparação entre amostras de diferentes usuários.

3.4. Classificação

A probabilidade de um usuário gerar características idênticas ao realizar uma atividade é baixa, pois muitos dos fatores utilizados na biometria comportamental são produzidos de maneira inconsciente ou involuntária pelo corpo. Por isso, estabelecemos um limiar que define o valor mínimo de similaridade que a classificação deve atingir para ser considerado, de fato, o mesmo usuário. Implementamos essa verificação utilizando aprendizado de máquina, permitindo que o modelo encontre relações entre as características, compare cada parâmetro com as amostras de treino e determine o nível de proximidade entre elas [Müller and Guido 2016]. Para esse estudo, foram utilizados os seguintes algoritmos: *SVM* (*Support Vector Machine*), *RF* (*Random Forest*), *KNN* (*K-Nearest Neighbors*) e *RL* (*Regressão Logística*). Esses algoritmos foram escolhidos com base nas preferências dos autores na literatura [Grzenda et al. 2023, Salmeron-Majadas et al. 2018, Singh et al. 2020].

4. Resultados Experimentais

Para evidenciar a eficiência do método, essa seção apresenta os resultados de um experimento realizado simulando uma situação cotidiana de transação online. Um *website* foi desenvolvido para coletar os dados necessários e alimentar o banco de dados criado pelos autores para conduzir os estudos. Todas as alterações foram limitadas ao *backend* e o monitoramento foi restrito apenas à tela usada para a coleta, prezando pela privacidade.

Solicitamos que 13 voluntários completassem 100 sessões cada. Cada voluntário dedicou, em média, 10 minutos por dia durante 30 dias para a coleta de dados. Como resultado, o banco de dados possui 1.300 amostras disponíveis para estudo. O banco de dados, a interface, a descrição das características e todo o código desenvolvido neste estudo estão disponíveis gratuitamente em um repositório público no GitHub[1].

Os modelos de aprendizado de máquina foram treinados com 70% das amostras de cada usuário, enquanto os outros 30% foram utilizados para testar a eficiência dos modelos. Foram realizados 30 testes com diferentes combinações de conjuntos de treino e teste, e a média dos resultados foi calculada para aumentar a precisão das métricas. Obtemos bons resultados com todos os algoritmos, e *Random Forest* se destaca em comparação aos demais, conforme mostrado na Tabela 2.

Classificador	**Acurácia**	**FRR**	**FAR**	**F1-Score**	**Validação Cruzada**	**Desvio Padrão**
RF	99,74%	0,26%	0,02%	99,74%	99,73%	0,30%
SVM	99,54%	0,46%	0,04%	99,54%	99,31%	0,58%
KNN	99,15%	0,85%	0,07%	99,14%	98,97%	0,66%
RL	99,48%	0,52%	0,04%	99,48%	99,29%	0,62%

Tabela 2. Métricas do desempenho dos classificadores.

Os resultados dos testes dos modelos são satisfatórios, mas não são tão bons quando aplicados a novas amostras. O algoritmo consegue distinguir o usuário legítimo dos demais com uma diferença considerável na acurácia. No entanto, na identificação do próprio usuário legítimo oscila entre 40% e 80%, indicando uma variabilidade significativa no desempenho. Segundo a literatura, outros autores que alcançaram métricas igualmente boas utilizaram entre 700 e 3.000 amostras provenientes de 19 a 200 participantes, o que pode explicar a discrepância entre bons resultados nos testes e os resultados medianos com novas amostras. Apesar das limitações, os resultados obtidos são notáveis, considerando que o trabalho está em fase inicial de desenvolvimento e aprimoramento. Há muito a ser melhorado nos dados, nas características extraídas e nos modelos, mas os resultados iniciais são promissores e têm potencial de melhoria significativa.

5. Conclusão e Trabalho Futuro

Considerando a ascensão da biometria comportamental na área de segurança cibernética, o objetivo deste trabalho foi contribuir com uma nova abordagem para a análise biométrica, envolvendo múltiplos identificadores. Apresentamos um método de análise biométrica para o reconhecimento de usuários, cujos experimentos iniciais demonstraram eficiência e resultados promissores. Para trabalhos futuros, pretendemos aumentar significativamente a quantidade de voluntários e amostras coletadas, visando obter uma base de dados mais robusta e melhorar a precisão das análises. Além disso, planejamos adicionar o *smartwatch* como um dos dispositivos monitorados, combinando seus dados com os do teclado e do mouse. Isso permitirá aumentar o número de características analisadas, proporcionando uma representação mais completa do usuário legítimo e, consequentemente, gerando resultados mais precisos e confiáveis. Esperamos que essas melhorias resultem em experimentos que gerem resultados ainda mais precisos e satisfatórios.

[1]https://github.com/SharpShards/autenticacao-de-sistemas-baseados-em-biometria-comportamental.git

Referências

Agarwal, N., Danielsen, N. F., Gravdal, P. K., and Bours, P. (2022). Contract cheat detection using biometric keystroke dynamics. In *2022 20th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pages 15–21.

Dias, T., Vitorino, J., Maia, E., Sousa, O., and Praça, I. (2023). Keyrecs: A keystroke dynamics and typing pattern recognition dataset. *Data in Brief*, 50:109509.

Grzenda, M., Kaźmierczak, S., Luckner, M., Borowik, G., and Mańdziuk, J. (2023). Evaluation of machine learning methods for impostor detection in web applications. *Expert Systems with Applications*, 231:120736.

Khan, S., Devlen, C., Manno, M., and Hou, D. (2024). Mouse dynamics behavioral biometrics: A survey. *ACM Computing Surveys*, 56.

Müller, A. C. and Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. "O'Reilly Media, Inc.".

Nnamoko, N., Barrowclough, J., Liptrott, M., and Korkontzelos, I. (2022). A behaviour biometrics dataset for user identification and authentication. *Data in Brief*, 45:108728.

Salmeron-Majadas, S., Baker, R. S., Santos, O. C., and Boticario, J. G. (2018). A machine learning approach to leverage individual keyboard and mouse interaction behavior from multiple users in real-world learning scenarios. *IEEE Access*, 6:39154–39179.

Singh, S., Inamdar, A., Kore, A., and Pawar, A. (2020). Analysis of algorithms for user authentication using keystroke dynamics. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, pages 0337–0341.

Çevik, N., Akleylek, S., and Koç, K. Y. (2021). Keystroke dynamics based authentication system. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 644–649.