**Projeto de Pesquisa e Planejamento de Atividades**

|  |  |  |
| --- | --- | --- |
| **Aluno**: Júlio César de Sousa Cabral | | **Data início curso**: 04/10/2022 |
| **Orientador**: Thiago Gentil Ramires | | **Defesa em:**  04/2024 |
| **Curso**: MBA Data Science e Analytics | **Modalidade**: Distância | Turma: 222 |

1. **Título do projeto** *(Inicial)*

**Autenticação biométrica contínua com dinâmica de toque em aplicação bancária de dispositivo móvel**

1. **Introdução**

O uso de aplicativos instalados em dispositivos móveis tem sido um desafio para a área de segurança da informação. A sensibilidade de sistemas específicos como o internet banking demanda contínuo aperfeiçoamento da verificação do usuário, sobretudo com a crescente disponibilidade de acesso aos dados biométricos coletados por sensores dos dispositivos.

O emprego da biometria na identificação de usuários de sistemas tem se mostrado eficiente em diversas aplicações. Essa tecnologia se fundamenta na particularidade das características do indivíduo, sejam fisiológicas ou comportamentais. A biometria fisiológica é medida diretamente de parte do corpo humano, como o reconhecimento de leitura facial, de impressão digital e de íris. Por outro lado, a biometria comportamental é obtida de forma indireta, passiva, com a obtenção de dados peculiares do comportamento durante a ação de um usuário (Nanavati et al, 2002).

Os sistemas biométricos são classificados em sistemas de identificação ou de verificação (Nanavati et al, 2002). A identificação envolve a determinação da identidade de uma pessoa a partir de um banco de dados sem o prévio conhecimento de qual pessoa se trata. Dessa forma, o processo de validação precisa percorrer toda a base de dados em busca de uma correspondência. O sistema de verificação, por sua vez, almeja identificar a correspondência de uma pessoa com a identidade de um determinado indivíduo já conhecido. A rigor, esse tipo de comparação 1:1 (um para um) relacionando um nome de usuário a dados é chamado de autenticação (Nanavati et al, 2002). Em dispositivos móveis, o conceito de autenticação contínua se refere à certificação da identidade de um usuário durante a navegação no sistema em uma sessão inteira (Georgiev et al. 2022a).

Na biometria comportamental, o termo genérico dinâmica de toque (touch dynamics) compreende a digitação de teclas (keystroke dynamics) e o toque de gesto (touch gesture). A tecnologia da dinâmica de digitação utiliza a maneira distinta de uma pessoa digitar padrões para realização de verificação mediante dados do intervalo de tempo que envolve o acionamento das teclas (Nanavati et al, 2002). Em contextos de digitação mais prolongada, também é possível avaliar parâmetros como a frequência de caracteres, uso de símbolos especiais, quantidade de acionamento de tecla de conserto de erro digitado (Kolakowsa, 2020). Em sistemas de autentificação por pressionamento de teclas, oito caracteres do nome do usuário e da senha são o tamanho mínimo para de dados coletados (Nanavati et al, 2002).

A dinâmica de gesto foi inicialmente inspirada no movimento do mouse dos computadores de mesa (Georgiev, 2022a) e no contexto da mobilidade refere-se ao processo de medir e acessar o ritmo do toque humano nas telas de dispositivos. Os dados das características do toque como tamanho, pressão e coordenadas espaciais podem ser considerados uma assinatura digital do usuário (Teh et al, 2016). Dados de movimento coletados por sensores como acelerômetro, giroscópio e magnetômetro também podem ser considerados características particulares adicionais na entrada de construção de modelos (Kolakowsa et al, 2020).

A viabilidade da autenticação contínua em aplicação em dispositivos móveis por meio de biometria comportamental tem sido objeto de diversos estudos (Vaishnav et al, 2022). Esses trabalhos consideram os dados de um ou mais sensores como variáveis explicativas para o fenômeno da classificação de um usuário com diferentes metodologias. No contexto bancário, uma notável proposta com as modalidades dinâmica de pressionamento de tecla e movimentos do telefone teve como princípio a maneira o usuário a informa uma senha padronizada de oito dígitos (Buriro et al, 2017). Foram usados os classificadores Naive Bayes, Rede Neural e Floresta Aleatória (Random Forrest) em situações com resultados melhores à medida que a amostra de treinamento crescia.

Outra proposta de autenticação contínua em aplicação bancária considerou os dados de texto digitados nesse contexto limitados a ponto de serem desprezados (Incel et al, 2021). Esse sistema coletou não apenas os dados dos gestos de toque na tela, mas também do acelerômetro, giroscópio e magnetômetro. O estudo se valeu apenas dos eventos de rolagem por serem os mais comuns no aplicativo original. Os participantes realizaram cinco ações mais frequentes no aplicativo modificado em três posturas, totalizando 15 sessões por usuário.

Os dois estudos citados voltados para autenticação de sistemas bancários, com dinâmica de digitação e de gestos fornecem direções conceituais, metodológicas e de desempenho para uma nova abordagem que considera os dados de ambos procedimentos. O uso de senha alfanumérica ainda é um meio comum de autenticação de usuários em aplicativos de bancos no Brasil, mesmo com a disponibilidade de outras modalidades. Embora a quantidade de caracteres possa variar conforme a instituição, uma avaliação em condições restritas pode ajudar na compreensão do desempenho de um sistema de segurança que recebe dados oriundos da digitação e a da dos gestos.

Espera-se que um usuário intruso em um sistema se comporte de maneira diferente do usual. A existência de dados modelados considerados normais faz que uma anomalia seja reconhecida como um desvio suficiente desse padrão (Aggarwal, 2017). Virtualmente, todos os algoritmos de detecção de anomalia criam um modelo do padrão normal dos dados e computa uma pontuação (score) do desvio de um determinado dado comparado. A avaliação do modelo é feita pela qualidade do ajuste entre o dado e o modelo. Dessa forma é possível identificar qual metodologia analisada apresenta melhor desempenho final.

1. **Objetivo**

A partir das abordagens em destaque (Buriro et al, 2017; Incel et al, 2021), almeja-se identificar se as duas abordagens atuando em conjunto, em uma mesma aplicação, podem melhorar o desempenho final. Pretende-se comparar os resultados com os dois estudos anteriores destacados. Além disso, busca-se constatar qual técnica de detecção da anomalia apresenta melhor desemepnho, bem como verificar quais variáveis coletadas pelos sensores impactam o resultado das avaliações.

1. **Material e Métodos**

Um aplicativo que simula ações de um hipotético banco será desenvolvido na plataforma Android. Esse sistema operacional foi escolhido por questão de facilidade de desenvolvimento e de ser o mais comum em celulares brasileiros. Inicialmente o armazenamento dos dados do treinamento será local, no próprio celular, enquanto na fase final de testes a aplicação acessará o modelo treinado disponível na rede.

Devido à dificuldade de disponibilidade dos dados específicos a serem estudados, será preciso realizar pesquisa de campo para coleta das informações de uso do aparelho. Serão seguidas as recomendações extraídas da análise de 30 trabalhos anteriores em sistemas de autenticação baseados em toque (Georgiev et al, 2022b):

a) Uso de apenas um aparelho celular na coleta de dados. Essa limitação ajuda a eliminar variações de resultados provenientes das diferenças entre os aparelhos. Ainda, de maneira geral, um invasor usa o mesmo celular da vítima em um ataque.

b) Amostra mínima de 40 usuários distintos. Toda a coleta será realizada em ambiente corporativo de um grande banco e locais públicos como shopping centers. Por questão de viabilidade da pesquisa, apenas parte do grupo inicial de entrevistados se submeterá posteriormente à fase de comprovação de verdadeiro positivo. Estes participantes possuem disponibilidade de segunda contribuição com a pesquisa.

c) Participantes desconhecidos emulam ataque invasor. O teste final de usuários que buscam efetuar autenticação em perfil diferente não fará parte da base de dados do treinamento para emulação de ataque mais próximo da realidade.

d) Contiguidade dos dados de treinamento. As ações dos usuários de celulares do mundo real são sequenciais ao longo do tempo. A escolha aleatória de uma amostra coletada pode trazer viés diferente do funcionamento dos sistemas reais. Além disso, haverá momento inicial de cadastro do usuário com senha predefinida seguido de navegação no aplicativo, para promover maior familiaridade com as ações que serão requisitadas.

A postura do usuário será controlada durante a pesquisa, pois impacta diretamente nos resultados (Syed et al, 2019). A postura pode ser definida como a combinação da orientação do aparelho (retrato ou paisagem) e a posição de uso (seguro em mão ou sobre uma mesa) (Syed et al, 2019). Para facilitar a aceitação dos participantes da pesquisa, o entrevistado utilizará o celular sentado, sem, entretanto, o suporte da mesa. A tela será fixada na orientação retrato que espelha o funcionamento dos aplicativos bancários.

Os sistemas de dinâmica de digitação costumam requerer longo treinamento com repetição de digitação de texto definido por aproximadamente 15 vezes (Nanavati et al, 2002), o que pode ser bastante desencorajador para um participante de pesquisa. Para contornar esse problema, as próprias operações de digitação de senha e valores das transações bancárias coletarão os dados das características de digitação do usuário. As senhas e os valores das operações serão fixados para maximizar a repetição do movimento durante o treinamento. Serão definidos sete códigos de CPF a serem digitados na operação de transferência PIX. Dessa forma, são esperados oito inserções de usuário e senha e sete de chaves da transação.

As seguintes ações no aplicativo serão requisitadas:

a) Cadastro de usuário no sistema. Nenhum dado será coletado pelos sensores nesse momento inicial.

b) Conectar ao aplicativo. A partir daqui os dados dos sensores são efetivamente armazenados. O usuário realiza o primeiro acesso com a informação do nome do usuário e senha. A cada conclusão da tarefa de movimentação PIX na conta, esse mesmo processo é evocado emulando uma autenticação bancária.

c) Navegação na sessão. Serão realizadas tarefas de sete transferências de valor a destinatários diferentes com chave PIX o código do CPF.

Uma folha impressa com as informações de cada uma das ações será disponibilizada para facilitar o entendimento do procedimento. O pesquisador lerá em alta voz cada ação da etapa e poderá ajudar realizar algum esclarecimento sobre a tarefa com o mínimo de interferência.

O sistema operacional Android fornece aos desenvolvedores API dos eventos disparados durante o uso com os dados brutos dos sensores. Esses dados podem ser classificados como temporais, espaciais e de movimento (Teh et al, 2016). Almeja-se, então, a partir dos eventos de digitação da senha, de CPF e de valores monetários, obter dados temporais e espaciais das informações:

a) Duração de digitação

b) Duração de tempo entre digitação

c) Tempo de pressão da tela

d) Área de pressão da tela

e) Velocidade de digitação na tela

Na prática, os valores temporais são calculados com a diferença do carimbo de data/hora com as ações do início (down) e conclusão do pressionamento (up) de uma tecla. O intervalo entre ações de teclas também é considerado importante variável a ser analisada. Também será seguida a criação de variáveis com medidas média, desvio padrão, mínimo, máximo, percentil, distância euclidiana, variância, curtose e assimetria com dados temporais.

De forma semelhante, os dados disparados pelos eventos do gesto na tela também possuem características temporais, com a diferença de também apresentarem coordenadas espaciais em duas dimensões, área e pressão do toque do dedo. O enfoque da abordagem da pesquisa será o movimento de rolagem de tela que alcança a opção do botão de transação do PIX. Inicialmente os dados serão coletados na frequência de 100 Hz à semelhança do estudo similar (Incel, 2021).

Adicionalmente, os dados tridimensionais de movimento do acelerômetro, giroscópio e magnetômetro também serão coletados pela API com a criação de vetor com a magnitude, ou seja, a raiz da soma dos quadrados dos valores de cada dimensão. Essa prática tem eficácia comprovada por estudos anteriores (Buriro, 2016).

Após a extração de dados, haverá fusão de todas as variáveis, com posterior seleção das características mais importantes e formação de modelo (template) em banco de dados. Serão implementados algoritmos comuns na análise de dinâmica de toque: Regressão Logística, Naive Bayes, K-Vizinhos mais Próximos (K-Nearest Neighbors) , Árvore de Decisão, Floresta Aleatória (Random Forest ) e Máquina de Vetores de Supoert (SVM).

As métricas utilizadas na avaliação do desempenho são as presentes em sistemas de biometria: a) Verdadeiros Positivos, b) Falsos Positivos, c) Falsos Negativos, d) Falsos Positivos, e) Acurácia, f) Proporção de erro equilibrada (EER, equal error rate). De fato, o EER foi a medida mais encontrada nos trabalhos pesquisados, embora não seja muito aplicável a sistemas reais (Nanavati, 2002).

1. **Resultados Esperados**

Espera-se superar tanto o desempenho do sistema Dakota que obteve média 90% de verdadeiros positivos e 3,5% de proporção de erro equilibrada, como a do trabalho com inserção de senha que alcançou taxa máxima de 96%. Concretamente, busca-se comprovar a evidência que a integração entre as abordagens de dinâmica de digitação e gesto produzem predição com maior acurácia (Wang et al, 2020).

De forma adicional, diante da dificuldade de acesso a dados da área de pesquisa, sobretudo em língua portuguesa, o banco de dados, todo o código que envolve a pesquisa será disponibilizado para a comunidade científica.

1. **Cronograma de Atividades**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atividades planejadas** | **Mês** | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| Desenvolvimento de aplicativo | **X** | X |  |  |  |  |  |  |  |  |
| Pesquisa de campo |  | **X** | **X** |  |  |  |  |  |  |  |
| Análise de dados e desenvolvimento de modelos |  |  |  | **X** | **x** |  |  |  |  |  |
| Entrega de resultados |  |  |  |  | **x** |  |  |  |  |  |
| Correções |  |  |  |  | **x** | **x** |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| Entrega do TCC |  |  |  |  |  | **x** |  |  |  |  |
| Entrega da apresentação |  |  |  |  |  |  | **X** |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

Projeto de Pesquisa; Resultados Preliminares; Entrega do Trabalho de Conclusão de Curso; Entrega da Apresentação da Defesa

1. **Referências Bibliográficas**

Aggarwal, C. C.. 2017. Outlier Analysis. 2ed. Springer, Yorktown Heights, New York USA.

Navavati, S.; Thieme, M.; Navavati, R. 2002. John Wiley & Sons, Inc . New York, USA

Raja L.; Vaishnav P.; Kaushik M. 2022. Design algorithm for continous autehntication on smartphone through keystroke dynamics and touch dynamics. Indian Journal of Computer Science and Engineering 13:444-455.

Kołakowska A.; Szwoch W.; Szwoch M. 2020. A Review of Emotion Recognition Methods Based on Data Acquired via Smartphone Sensors. Sensors 21: 6367.

Georgiev M.; Eberz S.; Martinovic I. 2022a. Techniques for Continuous Touch-Based Authentication Modeling. ArXiv:2207.12140. https://doi.org/10.48550/ARXIV.2207.12140.

Georgiev M.; Eberz S.; Turner H.; Lovisotto G.; Martinovic I. 2022b.Common evaluation pitfalls in touch-based authentication systems. In Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security. 1049--1063.

Incel, Ö.D.; Günay, S.; Akan, Y.; Barlas, Y.; Basar, O.E.; Alptekin, G.I.; Isbilen, M., 2021. DAKOTA: Sensor and Touch Screen-Based Continuous Authentication on a Mobile Banking Application 9:38943- 38959.

Zahid Syed, Jordan Helmick, Sean Banerjee, and Bojan Cukic. 2019. Touch gesture-based authentication on mobile devices: The effects of user posture, device size, configuration, and inter-session variability. Journal of Systems and Software 149 (2019), 158–173.

Teh, S. P.; Zhang, N.; Teoh, A. B.; Chen, K.A. 2016 . A survey on touch dynamics authentication in mobile devices. Computers & Security, 59:210-235.

Buriro, A.;Crispo, B.; DelFrari, F.; Wrona, K. 2016. Hold & Sign: A Novel Behavioral Biometrics for Smartphone User Authentication. IEE Security and Privacy Workshops. 276-285

A. Buriro, B. Crispo, F. DelFrari, and K. Wrona, “Touchstroke: Smartphone user authentication based on touch-typing biometrics,” in New Trends in Image Analysis and Processing–ICIAP 2015 Workshops. Springer, 2015, pp. 27–34.

Wang K.; Zhou, K.; Zhang, D.; Liu, Z.; Lim, J., 2020.What is More Important for Touch Dynamics based Mobile User Authentication? Twenty-Third Pacific Asia Conference on Information Systems, Dubai.