Edition: Vol. 03 | Nº. 01 | (2026)
Publication: 22/01/2026
DOI: https://doi.org/10.70579/pl.v3i1.130

Modelos explicáveis de aprendizado de máquina para detecção inteligente de comportamentos anômalos em ambientes zero-trust

Explainable machine learning models for intelligent detection of anomalous behavior in zero-trust environments

Modelos de aprendizaje automático explicables para la detección inteligente de comportamiento anómalo en entornos de confianza cero

Gabriel V Boechat
WB Educação
Pós graduação em Investigação Digital
Brasília - Distrito Federal, Brasil
gvboechat@gmail.com
http://lattes.cnpq.br/7484162897757683

RESUMO
A crescente complexidade dos ecossistemas digitais tem tensionado os modelos tradicionais de segurança da informação, baseados na noção de perímetro confiável. Nesse contexto, a Arquitetura Zero-Trust emerge como paradigma que redefine as práticas de controle de acesso, fundamentando-se no princípio de verificação contínua. Considerando que a efetividade desse modelo depende da identificação precisa de comportamentos anômalos, este estudo objetiva desenvolver e avaliar modelos de aprendizado de máquina integrados a técnicas de Inteligência Artificial Explicável (XAI) para a detecção de anomalias em ambientes Zero-Trust. Adota-se uma abordagem quantitativa e experimental, com uso de conjuntos de dados públicos amplamente reconhecidos na literatura de segurança computacional. São analisados algoritmos como Isolation Forest e One-Class SVM, combinados a métodos explicativos como SHAP e LIME. Os resultados indicam que a incorporação de XAI amplia significativamente a transparência das decisões automatizadas, sem comprometer o desempenho preditivo dos modelos. Conclui-se que modelos explicáveis contribuem para fortalecer a governança, a auditabilidade e a confiança operacional em arquiteturas Zero-Trust, oferecendo subsídios relevantes para aplicações práticas em ambientes organizacionais críticos.

Palavras-chave: zero-trust, detecção de anomalias, aprendizado de máquina, inteligência artificial explicável, segurança da informação.

.

Edition: Vol. 03 | Nº. 01 | (2026)
Publication: 22/01/2026
DOI: https://doi.org/10.70579/pl.v3i1.130

ABSTRACT
The increasing complexity of digital ecosystems has strained traditional information security models, based on the notion of a trusted perimeter. In this context, Zero-Trust Architecture emerges as a paradigm that redefines access control practices, based on the principle of continuous verification. Considering that the effectiveness of this model depends on the accurate identification of anomalous behaviors, this study aims to develop and evaluate machine learning models integrated with Explainable Artificial Intelligence (XAI) techniques for anomaly detection in Zero-Trust environments. A quantitative and experimental approach is adopted, using publicly available datasets widely recognized in the computer security literature. Algorithms such as Isolation Forest and One-Class SVM are analyzed, combined with explanatory methods such as SHAP and LIME. The results indicate that the incorporation of XAI significantly increases the transparency of automated decisions, without compromising the predictive performance of the models. It is concluded that explainable models contribute to strengthening governance, auditability, and operational trust in Zero-Trust architectures, offering relevant support for practical applications in critical organizational environments.

Keywords: zero-trust, anomaly detection, machine learning, explainable artificial intelligence, information security.

RESUMEN
La creciente complejidad de los ecosistemas digitales ha puesto a prueba los modelos tradicionales de seguridad de la información, basados en la noción de un perímetro de confianza. En este contexto, la Arquitectura de Confianza Cero surge como un paradigma que redefine las prácticas de control de acceso, basándose en el principio de verificación continua. Considerando que la eficacia de este modelo depende de la identificación precisa de comportamientos anómalos, este estudio busca desarrollar y evaluar modelos de aprendizaje automático integrados con técnicas de Inteligencia Artificial Explicable (IAE) para la detección de anomalías en entornos de Confianza Cero. Se adopta un enfoque cuantitativo y experimental, utilizando conjuntos de datos públicos ampliamente reconocidos en la literatura sobre seguridad informática. Se analizan algoritmos como Bosque de Aislamiento y SVM de Clase Única, combinados con métodos explicativos como SHAP y LIME. Los resultados indican que la incorporación de IAE aumenta significativamente la transparencia de las decisiones automatizadas, sin comprometer el rendimiento predictivo de los modelos. Se concluye que los modelos explicables contribuyen a fortalecer la gobernanza, la auditabilidad y la confianza operativa en las arquitecturas de Confianza Cero, ofreciendo un soporte relevante para aplicaciones prácticas en entornos organizacionales críticos.

Palabras clave: confianza cero, detección de anomalías, aprendizaje automático, inteligencia artificial explicable, seguridad de la información.

Edition: Vol. 03 | Nº. 01 | (2026)
Publication: 22/01/2026
DOI: https://doi.org/10.70579/pl.v3i1.130

1 INTRODUÇÃO

A intensificação dos processos de transformação digital alterou profundamente a forma como organizações produzem, armazenam e compartilham informações, ampliando a superfície de ataque e a complexidade dos ambientes computacionais. Tecnologias como computação em nuvem, mobilidade e integração de dispositivos heterogêneos fragilizaram a noção clássica de fronteira de rede, sobre a qual se estruturavam os modelos tradicionais de segurança da informação (Sommer; Paxson, 2010).

Nesse cenário, a Arquitetura Zero-Trust consolida-se como resposta conceitual e operacional às limitações dos modelos baseados em perímetro. Formalizada no âmbito do National Institute of Standards and Technology por meio do NIST SP 800-207, essa abordagem fundamenta-se no princípio de que nenhuma entidade deve ser presumida confiável, exigindo autenticação e autorização contínuas, baseadas em contexto, identidade e comportamento (Rose et al., 2020).

Entretanto, a efetividade do Zero-Trust depende fortemente da capacidade de identificar comportamentos anômalos que, muitas vezes, não se manifestam como ataques explícitos, mas como desvios sutis em padrões legítimos de uso. A literatura aponta o aprendizado de máquina como estratégia promissora para lidar com grandes volumes de dados e identificar tais desvios (Chandola; Banerjee; Kumar, 2009). Todavia, modelos de alto desempenho frequentemente operam como “caixas-pretas”, limitando a compreensão de suas decisões.

Essa opacidade torna-se particularmente problemática em ambientes Zero-Trust, nos quais decisões automatizadas impactam diretamente políticas de acesso, autenticação adaptativa e resposta a incidentes. Assim, emerge uma lacuna relevante na literatura: a integração sistemática entre detecção de anomalias baseada em aprendizado de máquina e técnicas de Inteligência Artificial Explicável (XAI), capazes de fornecer transparência, auditabilidade e suporte à governança.

Edition: Vol. 03 | Nº. 01 | (2026)
Publication: 22/01/2026
DOI: https://doi.org/10.70579/pl.v3i1.130

Diante desse contexto, este estudo tem como problema de pesquisa compreender de que maneira a incorporação de técnicas de XAI pode contribuir para aumentar a transparência e a confiabilidade operacional de modelos de detecção de comportamentos anômalos em arquiteturas Zero-Trust. O objetivo geral consiste em desenvolver e avaliar modelos explicáveis de aprendizado de máquina aplicados à detecção de anomalias nesses ambientes. Como objetivos específicos, busca-se: (i) analisar abordagens atuais de detecção de anomalias; (ii) implementar e comparar algoritmos de aprendizado de máquina; (iii) aplicar técnicas de explicabilidade; e (iv) avaliar desempenho e interpretabilidade dos modelos.

2 REFERENCIAL TEÓRICO

O referencial teórico deste estudo estrutura-se em três eixos conceituais centrais e interdependentes: (i) a Arquitetura Zero-Trust como paradigma emergente de segurança da informação; (ii) a detecção de comportamentos anômalos por meio de técnicas de aprendizado de máquina; e (iii) a Inteligência Artificial Explicável como resposta aos desafios de opacidade e governança dos modelos computacionais aplicados a contextos críticos. Essa organização permite sustentar analiticamente o problema de pesquisa e oferecer base teórica consistente para as escolhas metodológicas adotadas.

2.1 ARQUITETURA ZERO-TRUST E A REDEFINIÇÃO DOS MODELOS DE SEGURANÇA

Os modelos tradicionais de segurança da informação foram historicamente estruturados a partir da noção de perímetro, pressupondo que os ativos localizados no interior da rede organizacional seriam intrinsecamente confiáveis. Esse pressuposto mostrou-se progressivamente inadequado diante da expansão da computação em nuvem, da mobilidade de usuários, do uso intensivo de dispositivos heterogêneos e da crescente sofisticação das ameaças digitais, que passaram a explorar credenciais legítimas e

Edition: Vol. 03 | Nº. 01 | (2026)
Publication: 22/01/2026
DOI: https://doi.org/10.70579/pl.v3i1.130

movimentos laterais internos (Sommer; Paxson, 2010).

Nesse contexto, a Arquitetura Zero-Trust emerge como um paradigma que rompe com a lógica de confiança implícita. Sistematizada pelo National Institute of Standards and Technology no documento NIST SP 800-207, a abordagem Zero-Trust fundamenta-se no princípio de "nunca confiar, sempre verificar", estabelecendo que nenhuma entidade — usuário, dispositivo ou aplicação — deve receber acesso sem verificação contínua, independentemente de sua localização na rede (Rose et al., 2020).

A centralidade da identidade, a segmentação lógica dos recursos, a aplicação de políticas dinâmicas e a avaliação contínua do comportamento tornam-se elementos estruturantes desse modelo. Assim, a segurança deixa de ser um atributo estático e passa a operar como processo adaptativo, sensível ao contexto e às variações comportamentais. Nesse cenário, a capacidade de identificar desvios sutis em padrões legítimos de uso assume papel estratégico, pois ataques avançados tendem a se disfarçar como atividades normais do sistema.

2.2 DETECÇÃO DE COMPORTAMENTOS ANÔMALOS EM SEGURANÇA DA INFORMAÇÃO

A detecção de anomalias constitui um campo consolidado da literatura, voltado à identificação de padrões que divergem significativamente do comportamento esperado em determinado sistema. Chandola, Banerjee e Kumar (2009) propõem uma classificação abrangente das abordagens existentes, contemplando métodos estatísticos, técnicas baseadas em distância, modelos de clusterização e algoritmos de aprendizado de máquina supervisionados e não supervisionados.

No domínio da segurança da informação, métodos não supervisionados e semi-supervisionados ganharam destaque por sua capacidade de identificar eventos desconhecidos ou variações inéditas de ataques, sem depender exclusivamente de assinaturas previamente catalogadas. Algoritmos como Isolation Forest e One-Class Support Vector Machine são amplamente utilizados por explorarem a ideia de isolamento

Edition: Vol. 03 | Nº. 01 | (2026)
Publication: 22/01/2026
DOI: https://doi.org/10.70579/pl.v3i1.130

ou separação de observações raras em espaços de alta dimensionalidade.

Entretanto, a literatura aponta limitações relevantes dessas abordagens quando aplicadas a ambientes operacionais complexos. Sommer e Paxson (2010) destacam que modelos de aprendizado de máquina podem apresentar taxas elevadas de falsos positivos e dificuldades de interpretação, o que compromete sua adoção prática por equipes de segurança. Em arquiteturas Zero-Trust, esse problema é potencializado, pois decisões automatizadas influenciam diretamente políticas de acesso, autenticação adaptativa e resposta a incidentes.

Dessa forma, a eficácia da detecção de anomalias não pode ser avaliada apenas por métricas de desempenho, mas também pela capacidade do sistema de fornecer explicações compreensíveis, rastreáveis e alinhadas às necessidades de governança organizacional.

2.3 INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL E TRANSPARÊNCIA DOS MODELOS

A crescente adoção de modelos de aprendizado de máquina complexos intensificou o debate sobre a opacidade algorítmica, especialmente em contextos sensíveis como saúde, finanças e segurança da informação. A Inteligência Artificial Explicável (XAI) surge como campo de investigação voltado à compreensão, interpretação e transparência dos processos decisórios dos modelos computacionais.

Ribeiro, Singh e Guestrin (2016) propõem o LIME como método agnóstico de modelo, capaz de gerar explicações locais a partir da aproximação linear do comportamento do classificador em torno de uma instância específica. De forma complementar, Lundberg e Lee (2017) introduzem o SHAP, fundamentado na teoria dos valores de Shapley, oferecendo uma medida consistente da contribuição de cada atributo para a predição realizada.

Molnar (2022) enfatiza que a explicabilidade não deve ser compreendida apenas como recurso técnico adicional, mas como elemento essencial para a construção de sistemas confiáveis, auditáveis e socialmente responsáveis. Em ambientes Zero-Trust, a

Edition: Vol. 03 | Nº. 01 | (2026)
Publication: 22/01/2026
DOI: https://doi.org/10.70579/pl.v3i1.130

explicabilidade adquire relevância ampliada, pois decisões automatizadas precisam ser justificadas para operadores humanos, auditorias internas e conformidade regulatória.

Assim, a integração entre modelos de detecção de anomalias e técnicas de XAI não apenas amplia a compreensão dos resultados, mas também fortalece a confiança operacional, reduz a resistência organizacional à automação e contribui para práticas mais transparentes de segurança da informação.

3 METODOLOGIA

A pesquisa adota uma abordagem quantitativa e experimental, orientada à avaliação empírica de modelos de aprendizado de máquina aplicados à segurança da informação. O delineamento metodológico baseia-se em experimentos controlados, com uso de dados reais, reconhecidos na literatura científica.

Os conjuntos de dados selecionados correspondem a bases públicas amplamente utilizadas em estudos de detecção de intrusão, como CICIDS2017 e UNSW-NB15, por conterem registros de tráfego de rede associados a comportamentos legítimos e maliciosos. Esses dados passaram por etapas de pré-processamento, incluindo normalização, seleção de atributos relevantes e adequação conceitual aos princípios da Arquitetura Zero-Trust, como verificação contínua e centralidade da identidade.

Na etapa de modelagem, foram implementados algoritmos adequados à detecção de anomalias, com destaque para Isolation Forest e One-Class Support Vector Machine, selecionados por sua recorrência na literatura e aplicabilidade a cenários não supervisionados. O desempenho dos modelos foi avaliado por métricas como taxa de detecção e falsos positivos.

Para a dimensão explicativa, aplicaram-se técnicas de XAI, especialmente SHAP e LIME, visando identificar a contribuição das variáveis para as decisões dos modelos. As análises buscaram verificar se a explicabilidade compromete ou não a eficácia preditiva. Do ponto de vista ético, a pesquisa utilizou exclusivamente dados públicos, respeitando princípios de integridade científica, reprodutibilidade e transparência.

Edition: Vol. 03 | Nº. 01 | (2026)
Publication: 22/01/2026
DOI: https://doi.org/10.70579/pl.v3i1.130

4 RESULTADOS E DISCUSSÕES

Os resultados obtidos indicam que os modelos de aprendizado de máquina apresentaram desempenho satisfatório na identificação de comportamentos anômalos, confirmando achados prévios da literatura sobre a viabilidade de técnicas não supervisionadas em cenários de segurança computacional.

A aplicação de técnicas de XAI permitiu identificar padrões relevantes associados às decisões dos modelos, destacando atributos críticos relacionados à frequência de acessos, duração de sessões e variações abruptas no comportamento de usuários. Observou-se que a introdução da explicabilidade não resultou em degradação significativa do desempenho, mantendo taxas de detecção compatíveis com modelos não explicáveis.

Além disso, os resultados evidenciaram que a explicabilidade favorece a interpretação dos alertas gerados, oferecendo subsídios para analistas de segurança compreenderem os fatores subjacentes aos desvios identificados, o que se mostra especialmente relevante em ambientes Zero-Trust, nos quais decisões automatizadas impactam diretamente o fluxo operacional.

A análise dos resultados confirma que a integração entre aprendizado de máquina e técnicas de XAI constitui avanço relevante para a detecção de anomalias em arquiteturas Zero-Trust. Esses achados dialogam com estudos que apontam limitações práticas de modelos opacos em contextos críticos (Sommer; Paxson, 2010) e reforçam a necessidade de transparência destacada por Molnar (2022).

Em consonância com Ribeiro, Singh e Guestrin (2016) e Lundberg e Lee (2017), os resultados demonstram que métodos explicativos possibilitam compreender a influência das variáveis nos modelos, fortalecendo a confiança dos operadores e facilitando processos de auditoria e governança. No contexto Zero-Trust, essa

Edition: Vol. 03 | Nº. 01 | (2026)
Publication: 22/01/2026
DOI: https://doi.org/10.70579/pl.v3i1.130

característica revela-se estratégica, uma vez que decisões automatizadas exigem justificativas claras e rastreáveis.

Apesar das contribuições, o estudo apresenta limitações relacionadas à dependência de datasets públicos, que podem não capturar integralmente a complexidade de ambientes organizacionais reais. Pesquisas futuras podem explorar dados corporativos, modelos baseados em grafos e avaliações em ambientes produtivos, ampliando a validade externa dos resultados.

5 CONCLUSÃO

O presente estudo demonstrou que a integração entre modelos de aprendizado de máquina e técnicas de Inteligência Artificial Explicável constitui uma abordagem consistente e necessária para a detecção de comportamentos anômalos em ambientes fundamentados na Arquitetura Zero-Trust. Ao articular desempenho preditivo e interpretabilidade, a pesquisa evidenciou que é possível superar uma das principais limitações associadas ao uso de modelos complexos em segurança da informação: a opacidade decisória. Nesse sentido, os resultados reforçam que a explicabilidade não compromete a eficácia dos modelos, mas amplia sua utilidade prática, especialmente em contextos nos quais decisões automatizadas impactam diretamente políticas de acesso, autenticação contínua e resposta a incidentes.

Do ponto de vista teórico, o estudo contribui para o avanço da literatura ao integrar, de forma sistemática, três campos frequentemente tratados de maneira fragmentada: a Arquitetura Zero-Trust, a detecção de anomalias baseada em aprendizado de máquina e a Inteligência Artificial Explicável. Essa articulação permite compreender a segurança da informação não apenas como um problema técnico, mas como um sistema sociotécnico que exige transparência, rastreabilidade e alinhamento com práticas de governança. A pesquisa evidencia que a explicabilidade assume papel central na

Edition: Vol. 03 | Nº. 01 | (2026)
Publication: 22/01/2026
DOI: https://doi.org/10.70579/pl.v3i1.130

consolidação do Zero-Trust, ao oferecer subsídios interpretativos que fortalecem a confiança de analistas, gestores e auditores nos sistemas automatizados.

No plano metodológico, o trabalho reafirma a viabilidade do uso de algoritmos não supervisionados combinados a técnicas de XAI em cenários complexos e dinâmicos, como os ambientes Zero-Trust. A utilização de conjuntos de dados públicos e reconhecidos na literatura assegura reprodutibilidade e comparabilidade dos resultados, ao mesmo tempo em que evidencia limitações inerentes à ausência de dados organizacionais reais. Tais limitações, contudo, não invalidam os achados, mas indicam caminhos relevantes para investigações futuras, incluindo a aplicação dos modelos em ambientes produtivos, a incorporação de dados contextuais mais ricos e a exploração de abordagens baseadas em grafos ou aprendizado contínuo.

Por fim, conclui-se que modelos explicáveis de detecção de anomalias representam um avanço estratégico para a segurança da informação contemporânea, ao alinhar desempenho técnico, transparência algorítmica e exigências operacionais da Arquitetura Zero-Trust. Ao oferecer suporte à tomada de decisão informada e à governança de sistemas automatizados, esses modelos ampliam o potencial de adoção segura e responsável de soluções baseadas em inteligência artificial, contribuindo de maneira significativa para o fortalecimento da proteção de ambientes computacionais críticos.

REFERÊNCIAS

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: a survey**.** ***ACM Computing Surveys***, v. 41, n. 3, 2009.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. **Advances in Neural Information Processing Systems**, 2017.

MOLNAR, C. **Interpretable Machine Learning.** 2. ed. 2022.

Edition: Vol. 03 | Nº. 01 | (2026)
Publication: 22/01/2026
DOI: https://doi.org/10.70579/pl.v3i1.130

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why should I trust you?" Explaining the predictions of any classifier. **Proceedings of the 22nd ACM SIGKDD,** 2016.

ROSE, S. et al. **Zero Trust Architecture.** NIST Special Publication 800-207. Gaithersburg: NIST, 2020.

SOMMER, R.; PAXSON, V. Outside the closed world: On using machine learning for network intrusion detection. **IEEE Symposium on Security and Privacy**, 2010