ICS 35.240.01
CCS L70

团 体 标 准

T/COSOCC 012—2024

信息技术应用创新 基于人工智能的入侵检测产品技术要求

Information technology application innovation—Technical requirements for Artificial Intelligence-based intrusion detection products

2024－04－10 发布　　2024－04－10 实施

中国基本建设优化研究会　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国基本建设优化研究会提出并归口。

本文件起草单位：公安部第一研究所、云南省统计建模与数据分析重点实验室、安芯网盾（北京）科技有限公司、四川中电启明星信息技术有限公司、上海交通大学、北京浩瀚深度信息技术股份有限公司、重庆梦之想科技有限责任公司、北京信息科技大学、江苏蓝创智能科技股份有限公司、中企网络通信技术有限公司、拓尔思天行网安信息技术有限责任公司、北京亿中邮信息技术有限公司、广东盈世计算机科技有限公司、北京启明星辰信息安全技术有限公司、中电科申泰信息科技有限公司、北京中超伟业信息安全技术股份有限公司、北京网御星云信息技术有限公司、蓝象标准（北京）科技有限公司、嵩嘉标准化技术服务（北京）有限公司。

本文件主要起草人：李坤、唐年胜、姜向前、李欢欢、李高磊、吴晓春、郝艳、彭程、王新阳、赵纪元、黄青蓝、冯楠坪、王玉兰、范玲玲、卜令超、林延中、姚燕良、周昱、罗远哲、刘建军、段小莉、张德保、熊凡凡、姜冰、周紫晗、邱天、乔华阳。

引　言

在当前高质量发展阶段，智能化信息化的需求越来越强烈。特别是随着人工智能在各个领域的应用不断增加，网络安全面临的威胁也变得更加复杂。国际上，人工智能技术在网络安全领域的应用正不断发展，越来越多的企业和机构开始尝试使用利用人工智能技术来提高网络安全防御能力。因为国内安全厂商目前对人工智能重视程度不高，所以研究的不深，相对全球，没有成熟的产品。此外，随着ChatGPT等人工智能应用在各行各业的普及，网络安全威胁可能会进一步增加。

入侵检测是一项用于正式检测入侵行为的过程，其主要特征包括采集反常的使用模式、被利用的脆弱性及其类型、利用方式，以及入侵发生的时间和方式。随着攻防对抗的不断升级，攻击者使用的入侵手段变得日益复杂和隐蔽，传统的入侵检测方法对高级入侵手段的检测效果需要提升。所以人工智能技术在网络安全领域的应用具有巨大潜力。人工智能入侵检测方法通过分析网络数据流量、协议特征和主机行为等多种因素，能够自动识别入侵行为。这种方法通常对高级入侵手段具有较高的检出率，有效补充了传统的入侵检测方法。

目前，基于人工智能的入侵检测产品在性能和特点方面与传统方法存在差异和问题，例如需要更大的计算力和更高的成本等,但是目前尚无相关标准。本文件旨在规范人工智能入侵检测方法的检测效果，探索人工智能技术在网络安全领域的应用，提高网络安全保障能力，并推动行业对人工智能更加关注。通过标准固化先进技术、引领行业发展，为信息化建设和信息安全提供技术指导。

通过本文件的发布，可以提高人工智能入侵检测方法的准确性和可靠性，增强网络安全防御能力，探索人工智能技术在网络安全领域的应用，促进信息化建设和信息安全的发展。同时，提供对高级入侵手段的较高检出率，从而提高网络安全保障能力。本文件与GA/T 403.1相比，更加注重人工智能技术在网络安全领域的应用和发展，更加贴近实际应用需求。

信息技术应用创新 基于人工智能的入侵检测产品技术要求

1 范围

本文件规定了基于人工智能的入侵检测产品的总体架构、技术要求及系统配置要求。

本文件适用于指导基于人工智能的入侵检测产品的设计。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 25069 信息安全技术 术语

GB/T 37090 信息安全技术 病毒防治产品安全技术要求和测试评价方法

GA/T 403.1 信息安全技术 入侵检测产品安全技术要求 第1部分：网络型产品

GA/T 403.2 信息安全技术 入侵检测产品安全技术要求 第2部分：主机型产品

GA/T 1539 信息安全技术 网络病毒监控系统安全技术要求和测试评价方法

3 术语和定义

GB/T 25069、GB/T 37090、GA/T 403.1、GA/T 403.2、GA/T 1539界定的以及下列术语和定义适用于本文件。

3.1

恶意代码 malicious code

故意编制或设置的、对网络或系统会产生威胁或潜在威胁的计算机代码。

3.2

准确率 accuracy

对于给定的数据集，正确分类的样本数占总样本数的比率。

3.3

对抗性样本 adversarial examples

在数据集中通过故意添加细微的干扰所形成输入样本，受干扰之后的输入导致模型以高置信度给出错误的输出。

3.4

入侵检测系统 intrusion detection system

用于监测计算机网络和系统中潜在安全威胁和异常活动的安全工具，检测可能表明恶意攻击或未经授权的访问尝试的迹象，并向安全管理员或系统管理员发出警报。

4 缩略语

下列缩略语适用于本文件。

IDS：入侵检测系统（Intrusion Detection System）

IOA：攻击指标（Indicators of Attack）

IOC：攻陷指标（Indicators of Compromise）

5 总体架构

5.1 IDS

5.1.1 概述

基于网络流量检测原理和端点检测原理的人工智能入侵检测产品的分类见图1。

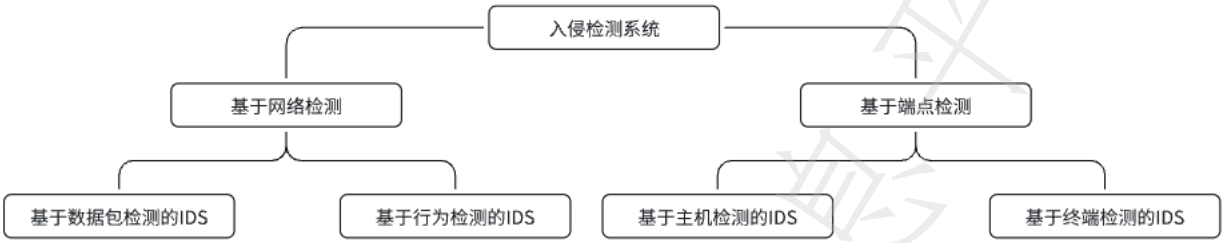

图1 人工智能入侵检测产品的分类

5.1.2 基于网络检测的 IDS

5.1.2.1 基于数据包检测的 IDS

数据包一般包含关通信双方、传输协议和数据内容的信息，通过预定义的数据包签名或规则来检测该网络段上发生的网络入侵，并可以对传输日志记录的网络通信相关信息进一步分析和检测。

5.1.2.2 基于行为检测的 IDS

根据非正常行为和使用计算机资源的情况检测入侵。通过对流量行为的检测学习和分类，识别非正常的活动和正常行为，从而检测潜在的入侵和威胁。

5.1.3 基于端点检测的 IDS

5.1.3.1 基于主机检测的 IDS

在每个要保护的主机上运行一个代理程序以检测入侵，一般只能检测该主机上发生的入侵。定义清楚不合法活动后，将这种安全策略转换成入侵检测规则。

5.1.3.2 基于终端检测的 IDS

在每个要保护的终端设备上运行一个代理程序以检测入侵，通过监测和分析终端设备上的操作系统和应用程序行为，以侦测潜在的入侵和恶意活动。

5.2 基于人工智能的 IDS

5.2.1 概述

基于人工智能的IDS可以利用机器学习、深度学习、大模型算法对网络流量、系统日志和其他相关数据进行分析和建模，以识别潜在的入侵行为：

a) 机器学习是让机器模拟人类的学习、思维、分析、判断的一种人工智能的深度学习算法；

b) 深度学习是机器学习的一个特定分支，它使用多层神经网络模型进行学习和训练；

c) 大模型是拥有大量参数和更大容量的机器学习或深度学习模型。

同时，基于人工智能的 IDS 需要满足传统 IDS 的通用要求，包括安全功能要求、性能要求、自身安全保护要求、环境适应性要求、安全保障要求。

5.2.2 人工智能应用于 IDS

人工智能技术运用智能化大数据分析算法，可以对数据进行判断，以区分哪些数据行为为正常行为或非正常行为。利用机器学习建立入侵检测模型，通过数据挖掘找出攻击间的关联性、分析出攻击行为

中的关联信息。通过训练机器学习模型，模型可以学习正常行为和非正常行为之间的差异，并在实时数据中识别出非正常行为活动，如异常或恶意活动。

在IDS中，机器学习算法被用来训练模型，自动提取和筛选有效的攻击特征，以解决传统方法中需要手动抓取检测数据特征的问题，提高入侵检测的效率。机器学习技术的发展促进了人工智能入侵检测研究的进步，基于监督和无监督机器学习的方法均在入侵检测领域有广泛的研究。人工智能应用于IDS的常用方法见图2。

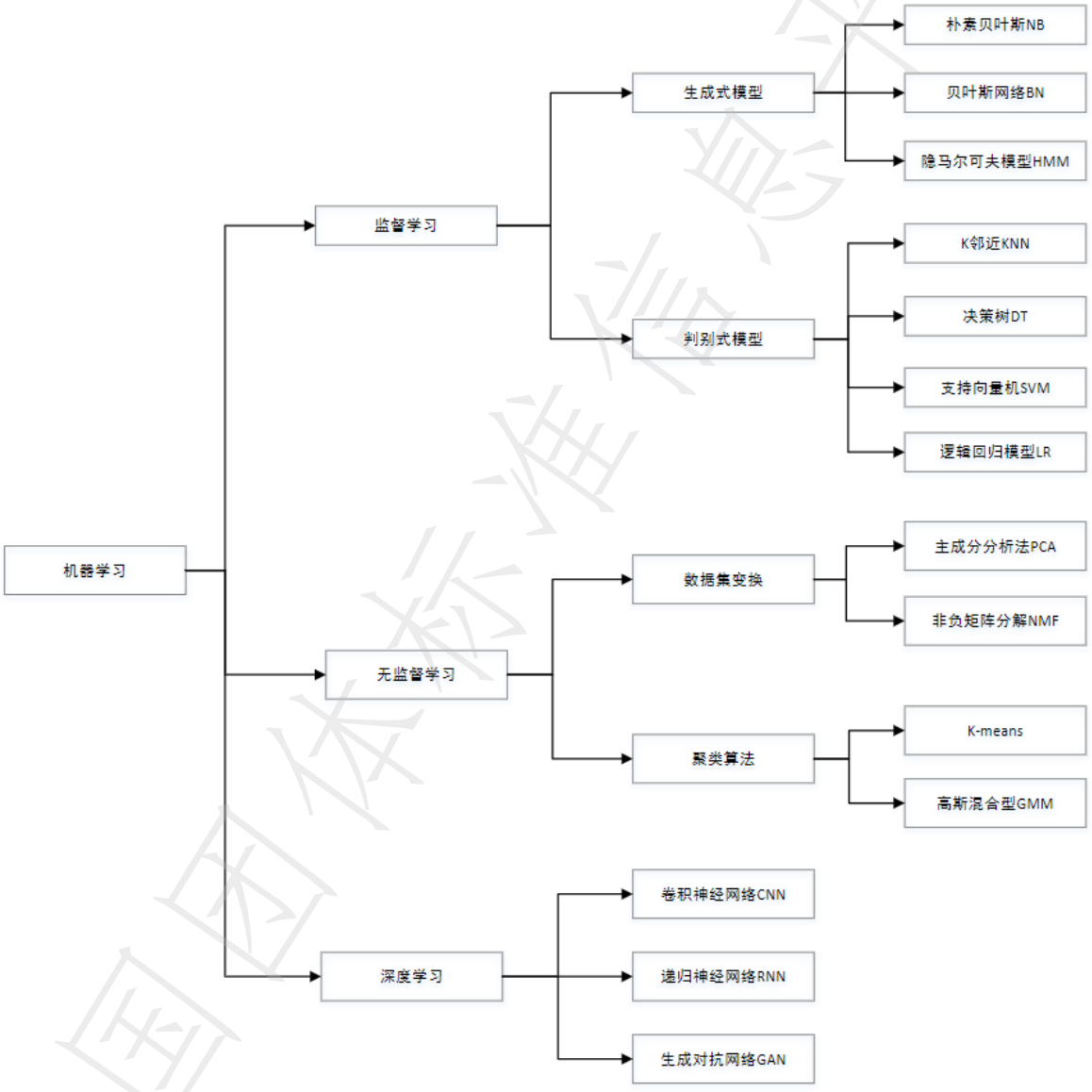

图2　人工智能应用于 IDS 的常用算法

5.2.3　基于网络检测的人工智能应用

5.2.3.1　基于数据包检测的人工智能应用

人工智能可以学习分析大规模的网络数据流，检测数据包中的非正常模式，如分布式拒绝服务（DDoS）攻击或恶意流量，从而快速识别潜在的网络入侵。

5.2.3.2 基于行为检测的人工智能应用

人工智能可以学习正常网络活动的行为模式，并自动检测不寻常的行为，例如异常的访问模式、大规模数据传输等，以识别潜在的威胁。

5.2.4 基于端点检测的人工智能应用

5.2.4.1 基于主机检测的人工智能应用

人工智能可以监测和分析主机操作系统和应用程序的行为，以检测恶意软件感染、未经授权的系统访问或不寻常的进程活动。

5.2.4.2 基于终端检测的人工智能应用

人工智能可以可以通过不同的学习模式检测终端上的异常文件、文件访问模式和文件传输，发现潜在的恶意文件等。

5.3 评测环境

人工智能入侵检测产品的评测环境见图3，是用于测试和评估该系统在检测和防御网络入侵方面的性能和效果的特定环境和设置。在检测过程中，应采用来自官方已披露漏洞的样本进行检测：如公共漏洞和暴露（CVE）、中国国家信息安全漏洞库（CNNVD）等。

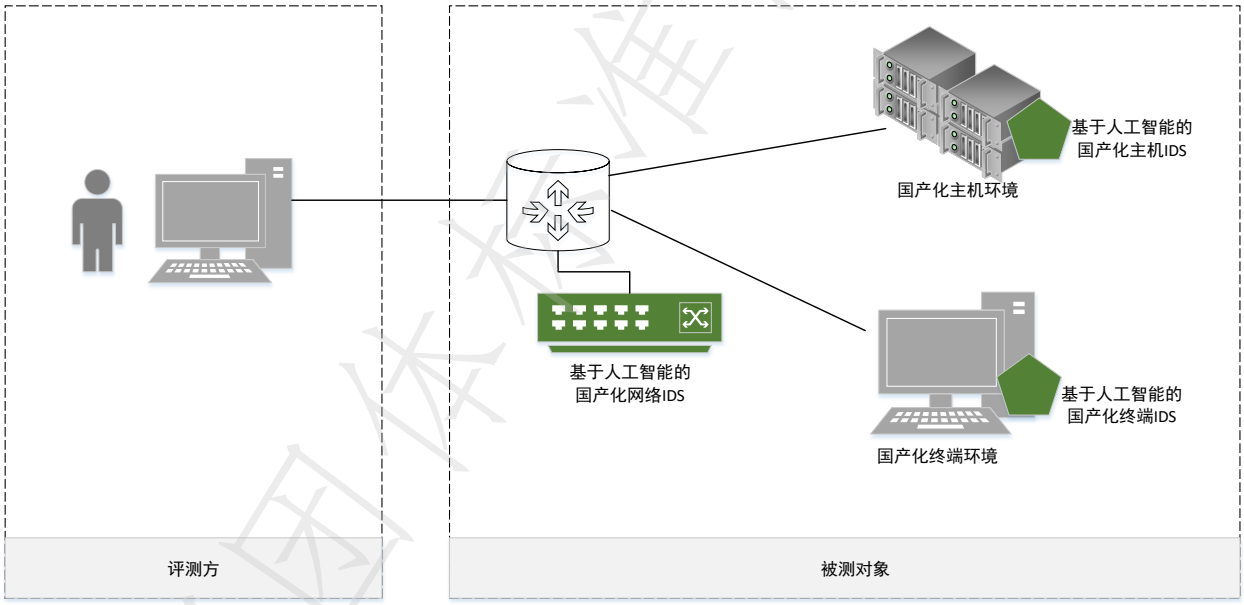

图3 人工智能入侵检测产品的评测环境

在测试过程中，要求关闭原有的非基于人工智能的检测能力，以确保检测的准确性和有效性，同时要求对关键检测过程进行记录，以明确该检测能力是基于人工智能的技术。

6 技术要求

6.1 人工智能算法的要求

6.1.1 高效性

算法应具备高效性能，能够在合理的时间内处理大规模的网络数据流，并及时识别和响应潜在的入侵行为。

6.1.2 准确性

算法应具备高准确率，能够有效地识别恶意代码、蠕虫、木马程序、Rootkit恶意程序、Bootkit恶意程序等各类入侵行为，减少误报和漏报的情况。

6.1.3 实时性

算法应具备实时性能，能够对实时产生的网络数据进行快速分析和检测，及时发现并应对新型的入侵攻击。

6.1.4 自适应性

算法应具备自适应能力，能够对不断变化的入侵行为进行自动学习和适应，及时更新模型以应对新的威胁。

6.2 入侵检测能力

6.2.1 IOC

6.2.1.1 基于网络检测的 IOC

人工智能入侵检测应使用人工智能算法检测以下基于网络检测的 IOC：

a) 暴力破解；
b) 木马后门；
c) 拒绝服务攻击；
d) 反弹 shell 等。

6.2.1.2 基于端点检测的 IOC

人工智能入侵检测应使用人工智能算法检测以下基于端点检测的 IOC：

a) 暴力破解
b) 非法提权；
c) 非法登录；
d) 恶意代码执行；恶意篡改；
e) 恶意文件、进程、应用；
f) 病毒文件；
g) 缓冲区溢出等。

6.2.2 IOA

6.2.2.1 基于网络检测的 IOA

人工智能入侵检测应使用人工智能算法检测以下基于网络检测的 IOA：

a) 端口扫描；
b) 可疑链接等。

6.2.2.2 基于端点检测的 IOA

人工智能入侵检测应使用人工智能算法检测以下基于端点检测的 IOA：

a) 端口扫描；
b) 异常登录；
c) 权限提升；
d) 重复的登录失败；
e) 服务启停、系统重启等。

6.3 人工智能方法的正确性

人工智能入侵检测应保证方法的正确性，确保其在实际应用中能够有效地识别和防御各类入侵行为。方法正确性的评估应包括但不限于以下方面：

a) 实验验证：通过在实际环境下的测试和验证，评估对已知入侵行为的检出率和准确率。
b) 对抗性样本测试：针对对抗性样本进行测试，验证对于针对性攻击和干扰的鲁棒性和稳定性。
c) 独立评估：通过独立的第三方机构或专家对方法进行评估，验证其有效性和可靠性。

6.4 人工智能目标函数

6.4.1 概述

人工智能入侵检测的目标函数旨在衡量模型对于检测网络中的入侵行为的性能和准确度。目标函数的设计应当能够区分正常网络流量和恶意入侵行为，并对恶意入侵行为进行有效的检测和分类。人工智能入侵检测方法目标函数应满足分类准确性、假阳性和假阴性控制、目标置信度、对抗性样本鲁棒性和训练数据覆盖性。

6.4.2 分类准确性

目标函数应能够准确地分类网络流量，将正常流量和恶意入侵流量区分开来。分类准确性是评估模型性能的重要指标，可以使用准确率、精确率、召回率等指标来衡量分类结果的准确程度。

6.4.3 假阳性和假阴性控制

入侵检测中，假阳性是指将正常流量错误地分类为恶意入侵，而假阴性是指将恶意入侵错误地分类为正常流量。目标函数应尽可能降低假阳性和假阴性的数量，以提高模型的准确性和可靠性。

6.4.4 目标置信度设置

目标函数可以要求模型输出针对每个样本的置信度或概率值，以表示模型对于该样本为恶意入侵的预测程度。目标置信度可以用于进一步的决策过程或阈值设置，以实现定制化的入侵检测策略。因此要求产品可以通过调整目标函数的置信度阈值来调整决策结果。

6.4.5 对抗性样本鲁棒性

入侵检测模型应具备一定程度的对抗性样本鲁棒性，即对于经过修改或攻击的输入样本仍能保持较高的检测性能。故目标函数应要求模型在面对对抗性样本时表现稳定，并具备抵抗常见攻击技术的能力。

6.4.6 训练数据覆盖性

训练数据的多样性和充分性对于模型的泛化能力和鲁棒性至关重要。目标函数要求模型在训练过程中能够充分利用不同类型、不同来源的训练数据，以覆盖不同的入侵行为和网络环境。

6.5 训练数据集

6.5.1 充分性

训练数据集应包含丰富多样的入侵行为样本，涵盖各种暴力破解、非法提权、非法登录、恶意代码、恶意文件、病毒文件、拒绝服务攻击、木马后门等不同类型的入侵攻击，以覆盖多个攻击场景和数据分布。

6.5.2 样本种类和数量

样本种类和数量的多少可以影响模型的泛化能力和准确性，数据集应具有足够的样本种类和数量，以覆盖各种入侵行为和正常行为。数据集规模需要大于1 TB，低于规模值则不能构成样本。

6.5.3 样本质量和准确性

训练数据集中的样本应具有高质量和准确的标签，其中包括正常行为和不同类型的入侵行为。标记可以是二分类的（正常行为或入侵行为）或多类别的（如具体的入侵类型），以确保模型能够正确学习和识别不同类别的行为。

6.5.4 样本平衡性

训练数据集应具备样本类别平衡，避免过度关注某些入侵类型，导致对其他入侵行为的检测效果不佳。

6.5.5 实时性

由于入侵行为不断演变和新型攻击的出现，入侵检测领域的训练数据集应具备一定的实时性和在线学习能力，及时反映最新的入侵行为和攻击模式，能够随着新的入侵行为的出现进行更新和补充。

6.5.6 防范措施

由于模型训练和测试数据之间的分布差异可能发生变化，因此数据集需要具备一定的防范措施。应持续检查数据的质量，包括缺失值、异常值等。此外，应采取数据清洗、填充和预处理措施，以确保输入数据集的准确性。

6.5.7 数据隐私和安全

入侵检测数据集可能包含真实的攻击行为和敏感信息，需要确保数据的隐私和安全，应采取适当的数据脱敏和保护措施，确保数据集不会被滥用或泄露。

6.6 对抗性样本

人工智能入侵检测方法的对抗性样本应包括下列内容。

a) 白盒方式生成的样本：指目标模型已知的情况下，利用梯度下降等方式生成对抗性样本。
b) 黑盒方式生成的样本：指目标模型未知的情况下，利用一个替代模型进行模型估计，针对替代模型使用白盒方式生成对抗性样本。
c) 指定目标生成的样本：指利用已有数据集中的样本，通过指定样本的方式生成对抗性样本。
d) 不指定目标生成的样本：指利用已有数据集中的样本，通过不指定样本（或使用全部样本）的方式生成对抗性样本。

6.7 环境数据

为有效地进行入侵检测和安全防护，人工智能入侵检测方法的实际运行环境数据应包括下列内容。

a) 干扰数据：指由于环境的复杂性所产生的非预期的真实数据，可能影响算法的可靠性。
b) 数据集分布迁移：算法通常假设训练数据样本和真实数据样本服从相同分布，但在算法实际使用中，数据集分布可能发生迁移，即真实数据集分布与训练数据集分布之间存在差异性。
c) 野值数据：指一些极端的观察值，在一组数据中可能有少数数据与其余的数据差别比较大，也称为异常观察值。
d) 正常行为数据：指合法用户或系统的正常操作行为数据。通过建立正常行为模型，可以与后续的行为进行比较，从而检测出异常或可疑行为。
e) 异常行为数据：指潜在的入侵或异常行为的数据，包括已知的入侵行为、恶意软件、网络攻击等。通过分析异常行为数据，可以识别和检测恶意活动和入侵行为。
f) 网络流量数据：包括网络数据包、网络连接记录等网络流量信息。
g) 主机日志数据：包括操作系统日志、应用程序日志等主机产生的日志信息。
h) 安全事件数据：包括已知的安全事件记录、恶意代码样本等已知的入侵行为信息。

6.8 功能要求

6.8.1 概述

人工智能入侵检测产品应具有有效性和可用性，具备实时监测、入侵检测、威胁情报分析、威胁告警和响应、日志记录和分析、信息可视化和溯源等功能。

6.8.2 实时监测

应能实时监测网络流量和主机行为，及时发现潜在的入侵行为。

6.8.3 入侵检测

应能准确识别和分类各类入侵行为，包括暴力破解、非法提权、非法登录、恶意代码、恶意文件、病毒文件、拒绝服务攻击、木马后门等。

6.8.4 威胁情报分析

应能分析和评估威胁情报，及时更新入侵检测规则和模型。

6.8.5 安全告警

应能在检测到告警时自动采取相应动作以发出安全告警。

6.8.6 告警方式

应能支持多种告警方式，包括屏幕实时提示、E-mail告警、声音告警等几种方式。

6.8.7 响应能力

应能根据相应威胁告警采取相应的防御措施，包括阻断网络连接、隔离受感染主机等。

6.8.8 日志记录和分析

应能记录和分析入侵检测过程中的日志信息，支持后续的安全事件调查和分析。

6.8.9 信息可视化和溯源

应能提供包括网络流量可视化、异常检测结果可视化、事件分析可视化等功能的可视化界面，能快速对入侵来源进行溯源。

6.9 性能指标

6.9.1 准确率

系统对入侵行为的准确识别率应不低于传统入侵检测产品，准确率大于90%。

6.9.2 误报率

系统误报入侵行为的比率应尽可能低，误报率应该低于10%。

6.9.3 检测响应时间

在足够的硬件资源和高效的数据处理速度下，系统对入侵行为的检测和响应时间应不超过5 s。

6.9.4 鲁棒性

系统对对抗性样本和新型入侵攻击应具备较强的抵抗能力。

6.9.5 泛化性

系统对新的、未知的入侵攻击仍能表现出良好的性能和识别能力。

7 系统配置要求

7.1 处理器架构适配

应适配X86、进阶精简指令集机器（ARM）、每秒能够执行的百万条指令的数量（MIPS）等多种处理器架构。

7.2 操作系统适配

应适配麒麟（KylinOS）、欧拉（EulerOS）、统信（UOS）等国产操作系统。

7.3 跨平台适配

系统应能够在不同的硬件平台和操作系统上运行，满足不同用户的部署需求。

7.4 网络环境适配

系统应能够适应不同网络环境的特点和要求，包括云环境、内外网等不同的网络拓扑。

7.5 系统集成适配

系统应能够与现有的网络安全设备和系统进行集成，提供全面的入侵检测和防御能力。

7.6 可定制化适配

系统应提供可定制化的功能和配置选项，以满足不同用户的特定需求和业务场景。

参 考 文 献

[1] GB/T 20275—2021 信息安全技术 网络入侵检测系统技术要求和测试评价方法
[2] GB/T 26269—2010 网络入侵检测系统技术要求
[3] GA/T 1539—2018 信息安全技术 网络病毒监控系统安全技术要求和测试评价方法