面向服务的AI安全技术研究与实践

吴小珍
杭州安恒信息安全技术有限公司　浙江　杭州　310000

摘　要：在全球数字化转型进程迅猛推进的当下，AI技术深度融入信息系统的各个角落，成为推动各行业创新发展的关键力量。然而，AI的广泛应用也使得安全问题愈发突出，严重制约着AI技术的进一步发展与应用。本文深度聚焦于面向服务的AI安全技术，全面且深入地探究其关键技术的原理、应用场景以及在实践中的具体应用，致力于提升AI系统的安全性、可靠性与稳定性，为信息系统的稳健运行提供坚实有力的保障。

关键词：面向服务；AI安全技术；信息系统；应用实践

引言

在信息技术日新月异的时代，AI凭借强大的数据分析与处理能力，在信息系统领域广泛应用，从互联网企业的智能客服提升服务体验，到金融机构利用AI进行风险评估辅助投资决策，再到医疗领域借助AI提高疾病诊断准确性，极大地推动了产业升级与变革。但AI技术在带来便利的同时，也面临严峻安全挑战，像2017年美国Equifax公司数据泄露事件，严重侵害用户隐私。此外，算法可能存在漏洞、模型易受对抗样本攻击、应用层面权限管理不当导致数据滥用。面向服务的AI安全技术以服务为核心，从数据全生命周期管理、模型稳健性保障等多维度构建防护体系，对推动AI在信息系统安全、稳定、可靠应用及各行业数字化转型意义重大。

1 AI 安全技术基础与理论

1.1 AI系统架构与安全风险

AI系统架构剖析：AI系统是一个复杂的体系，通常由数据层、算法层、模型层和应用层构成。数据层作为AI系统的基础，负责海量数据的收集、存储与预处理工作。数据来源广泛，包括传感器数据、用户行为数据、业务交易数据等。例如，在智能交通系统中，数据层会收集来自车辆传感器的行驶速度、位置信息，以及交通摄像头拍摄的路况视频数据等。这些原始数据往往存在噪声、缺失值等问题，需要经过清洗、去噪、归一化等预处理操作，才能为后续的分析训练提供高质量的数据支持。

算法层是AI系统的核心大脑，运用机器学习、深度学习等多种先进算法对数据进行深度分析与训练。机器学习算法如决策树、支持向量机等，通过对大量数据的学习，建立数据特征与目标之间的关系模型。深度学习算法则以神经网络为基础，通过构建多层神经元结构，自动提取数据的高级特征。以图像识别领域为例，卷积神经网络（CNN）在图像特征提取与分类任务中表现卓越。它通过卷积层、池化层和全连接层的组合，能够有效地提取图像中的边缘、纹理等特征，从而实现对不同物体的准确识别。

模型层是AI系统的智慧结晶，经过算法层的训练，生成并存储训练好的模型。这些模型是对数据规律的抽象表达，能够根据输入数据进行预测和决策。在语音识别系统中，训练好的声学模型和语言模型能够将用户的语音信号转换为文字信息。模型的性能和准确性直接影响到AI系统的应用效果，因此需要不断优化和更新模型。

应用层是AI系统与用户和业务场景的交互界面，将训练好的模型应用于实际业务中，为用户提供各种智能化服务。如安防监控中的人脸识别系统，通过应用层将图像识别模型与监控摄像头相结合，实时对监控画面中的人脸进行识别和比对，实现人员身份验证、考勤管理等功能。

安全风险类型：数据层面，数据泄露风险时刻威胁着AI系统的安全。黑客可能通过网络攻击手段，入侵数据库，获取敏感数据。数据篡改也是常见的风险之一，攻击者可能会修改数据的内容，导致AI系统基于错误的数据进行决策。在医疗数据领域，若患者的病历数据被篡改，可能会影响医生的诊断和治疗方案[1]。

算法层面，算法漏洞可能被恶意利用。例如，一些机器学习算法对数据的分布较为敏感，攻击者可以通过精心设计的数据样本，使算法产生偏差，从而操纵AI系统的决策结果。在金融风险评估中，若算法被攻击，可能会导致错误的风险评估，给金融机构带来巨大损失。

模型层面，对抗样本攻击是主要的安全风险。攻击者通过对原始样本添加微小的扰动，生成对抗样本，使模型产生错误的输出。在自动驾驶领域，若图像识别模型受到对抗样本攻击，可能会导致车辆对交通标志的误

判，引发严重的交通事故。

应用层面，权限管理不当会导致数据滥用。一些企业可能会过度收集用户数据，并在未经用户同意的情况下，将数据用于其他商业用途。在社交网络平台中，若用户的个人信息被滥用，可能会导致用户的隐私泄露和骚扰。

1.2 安全技术理论基础

密码学原理：在AI安全领域，密码学是保障数据安全的重要基石。密码学主要用于数据的加密传输与存储，确保数据在传输和存储过程中的机密性、完整性和不可抵赖性。对称加密算法如AES（高级加密标准），它使用相同的密钥对数据进行加密和解密，具有加密速度快、效率高的特点，广泛应用于大量数据的加密场景。在云存储中，用户的数据可以使用AES算法进行加密后存储在云端，防止数据被窃取。

非对称加密算法如RSA，它使用一对密钥，即公钥和私钥。公钥用于加密数据，私钥用于解密数据。RSA算法主要用于数字签名和密钥交换，确保数据的完整性和不可抵赖性。在电子合同签署中，发送方使用私钥对合同内容进行数字签名，接收方使用发送方的公钥进行验证，保证合同内容未被篡改且发送方无法否认签署行为。

访问控制理论：基于角色的访问控制（RBAC）在AI系统中得到了广泛应用。RBAC根据用户在系统中的角色分配相应的权限，不同角色具有不同的操作权限和数据访问权限。例如，在一个企业的信息系统中，管理员角色拥有最高权限，可以管理系统的配置、用户权限、数据备份等操作；普通员工角色只能访问和处理与自己工作相关的数据，如查看自己的工作任务、提交工作报告等。通过RBAC，可以有效地防止权限滥用，保障系统的安全。

2 面向服务的 AI 安全关键技术

2.1 数据安全技术

数据加密技术：同态加密技术是一种新兴的数据加密技术，它允许在密文上进行计算，而无需对数据进行解密。在联邦学习中，同态加密技术发挥着重要作用。联邦学习是一种分布式机器学习技术，多个参与方在不共享原始数据的情况下，共同训练一个模型。各参与方将本地数据加密后上传，在加密状态下进行模型训练。例如，在医疗领域，多家医院可以通过联邦学习共同训练疾病诊断模型，同时保护患者的隐私数据不被泄露。同态加密技术基于复杂的数学原理，如基于格密码理论构建加密体系，确保密文计算的准确性和安全性，相比传统加密技术，在保护数据隐私的同时不影响数据的分析利用效率。

数据脱敏技术：数据脱敏技术是对敏感数据进行变形处理，使其在保持原有数据特征和业务逻辑的前提下，无法直接识别出用户的真实身份信息。对于用户的身份证号，可以采用部分隐藏的方式，将中间几位数字替换为星号；对于手机号，可以将中间几位数字替换为其他字符[2]。数据脱敏技术在保证数据可用性的同时，有效地保护了用户的隐私。在实际应用中，数据脱敏技术还会根据不同行业需求，如金融行业对银行卡号脱敏，采用特定的脱敏规则，保证数据在测试、分析等场景下既能保留业务价值，又能防止隐私泄露。

2.2 模型安全技术

模型加固技术：对抗训练是一种有效的模型加固方法，通过让模型学习对抗样本，增强模型对攻击的抵抗力。在图像识别模型训练中，将正常样本和对抗样本同时输入模型进行训练，使模型能够识别出对抗样本的特征，从而在面对对抗攻击时能够准确识别。例如，在安防监控中的人脸识别系统，通过对抗训练，可以提高模型对恶意篡改图像的识别能力。在对抗训练过程中，通过动态调整对抗样本的生成策略，如改变扰动的幅度和方向，使模型不断适应复杂的攻击场景，提升模型的泛化性和鲁棒性。

模型水印技术：模型水印技术是为模型添加唯一的标识信息，用于版权保护和模型来源追踪。水印信息可以是一段特殊的代码或数据，嵌入到模型的参数中。当发现模型被非法使用时，可以通过提取水印信息追溯模型的来源。在一些商业AI模型中，模型水印技术可以保护企业的知识产权，防止模型被抄袭和滥用[3]。水印嵌入过程会考虑模型的性能影响，采用优化算法，如基于梯度的水印嵌入方法，在保证水印不可见性的同时，最小化对模型精度的干扰。

3 AI 安全技术在信息系统中的应用实践

3.1 金融信息系统应用

风险评估与防范：金融信息系统中，利用AI安全技术对金融交易数据进行加密存储和分析，能够有效识别异常交易行为。通过机器学习算法建立风险评估模型，对大量的历史交易数据进行学习，提取正常交易和异常交易的特征。实时监测交易风险，当发现一笔交易的特征与异常交易特征匹配时，及时发出预警信号，防止金融欺诈行为的发生。例如，在信用卡交易中，若发现一笔交易的金额、地点、消费频率等特征与持卡人的历史交易习惯不符，系统会自动发出预警，要求持卡人进行身份验证。

客户身份验证：采用生物识别技术结合加密算法，进行客户身份验证。人脸识别、指纹识别等生物识别技术具有唯一性和不可复制性，通过加密传输和比对，确保客户身份的真实性。在移动支付中，用户可以通过人脸识别或指纹识别进行支付验证，加密算法保证了生物特征数据在传输和存储过程中的安全，有效保障了金融交易的安全。

3.2 医疗信息系统应用

患者数据保护：医疗信息系统中，患者数据包含大量的敏感信息，如病历、诊断结果、基因数据等。采用区块链技术，实现医疗数据的分布式存储和加密共享。区块链具有去中心化、不可篡改、可追溯等特点，患者的数据被加密后存储在多个节点上，只有经过患者授权的医生才能访问相应的数据。患者可以通过区块链技术查看自己的数据访问记录，保护自己的隐私。

医疗决策支持系统安全：医疗决策支持系统中的AI模型对于医生的诊断和治疗决策具有重要的参考价值。对医疗决策支持系统的AI模型进行加固，防止模型被攻击导致错误诊断。通过模型水印技术，确保模型的可靠性和来源可追溯[4]。在疾病诊断中，若模型被攻击导致错误的诊断结果，医生可以通过水印技术追溯模型的来源，判断模型的可靠性。

4 AI 安全技术实践效果评估与挑战

4.1 实践效果评估

安全指标提升：通过在金融、医疗等信息系统中的应用，AI安全技术取得了显著的成效。数据泄露事件显著减少，模型的准确率和抗攻击能力得到了大幅提高。金融机构应用AI安全技术后，数据泄露事件降低了80%，异常交易识别准确率提升了25%，有效保护了客户的资金安全和隐私。

业务效率提升：安全技术的应用保障了信息系统的稳定运行，提高了业务处理效率。在医疗信息系统中，患者数据的快速、安全共享，使医生能够更及时地获取患者的完整病历信息，提高了诊断效率。例如，在远程医疗中，患者的病历数据可以通过安全的传输通道快速传输给专家，专家可以及时进行诊断和治疗建议。

4.2 面临挑战

技术复杂性：AI安全技术涉及密码学、机器学习、区块链等多个领域的知识，技术实现复杂，增加了开发和维护成本。量子计算技术的发展可能对现有加密算法构成威胁，需要不断更新加密技术，以适应新的安全挑战。例如，量子计算机可能会破解传统的RSA加密算法，需要研究新的抗量子加密算法。

法律法规不完善：目前AI安全相关的法律法规尚不完善，在数据隐私保护、责任界定等方面存在空白。不同国家和地区的法律法规存在差异，给跨国企业的AI安全技术应用带来了困难。在数据泄露事件中，责任界定不明确，导致受害者难以获得有效的赔偿。

结语

面向服务的AI安全技术在保障信息系统安全、推动AI技术在各行业的安全应用方面取得了显著的成效。通过数据安全技术、模型安全技术等关键技术的应用，有效地降低了AI系统的安全风险，提升了业务效率，为数字化发展提供了有力的支持。AI安全技术的发展仍面临着诸多挑战，技术复杂性和法律法规不完善等问题亟待解决。未来，需要加强跨学科的研究与合作，加大对AI安全技术的研发投入，不断创新和完善安全技术体系。同时，政府和相关机构应加快制定和完善AI安全相关的法律法规，明确各方的责任和义务，为AI安全技术的应用创造良好的法律环境。只有这样，才能推动AI安全技术在信息系统中更广泛、更深入地应用，为数字化时代的安全与发展筑牢坚实的防线。

参考文献

[1]叶菁.AI与云打造安全屏障共同构建云上智安新生活[N].通信信息报,2025-01-08(002).

[2]本报编辑部.360集团创始人周鸿祎:打造安全大模型,用AI重塑安全[N].中国计算机报,2025-01-06(005).DOI:10.28468/n.cnki.njsjb.2025.000001.

[3]曾懿辉,张虎,麦俊佳.基于AI与数字孪生技术的高空作业安全监测方法[J/OL].自动化技术与应用,1-6[2025-02-20].http://kns.cnki.net/kcms/detail/23.1474.TP.20241230.0939.076.html.

[4]孙亚楠.AI技术在建筑施工安全管理中的应用与伦理探究[J].山西建筑,2025,51(04):187-189+194.DOI:10.13719/j.cnki.1009-6825.2025.04.041.