欺诈的敌人：信用 card 欺诈检测中的可转移对抗攻击

Jan Lum Fok[a,b], Qingwen Zeng[a],Shiping Chen[c], Oscar Fawkes[b], Huaming Chen[a]
[a] The University of Sydney, Australia
[b] Constantinople, Australia
[c] CSIRO, Australia
Jan@cxnpl.com, {qingwen.zeng, huaming.chen}@sydney.edu.au, shiping.chen@data61.csiro.au, OscarF@cxnpl.com

摘要—信用卡欺诈检测（CCFD）是金融领域机器学习（ML）的一项关键应用，准确识别欺诈交易对于减轻财务损失至关重要。ML 模型在欺诈检测任务中已经证明了其有效性，特别是在表格数据集上。虽然对抗攻击已在计算机视觉和深度学习中得到广泛研究，但它们对 ML 模型的影响，特别是那些基于 CCFD 表格数据训练的模型，仍然很大程度上未被探索。这些潜在漏洞对金融行业的安全和稳定构成了重大威胁，尤其是在高价值交易中，损失可能相当大。为了解决这一差距，在本文中，我们提出一个全面框架来研究在不同情况下，CCFD ML 模型对抗扰动的鲁棒性。具体而言，基于梯度的攻击方法被纳入了信用卡交易表格数据中的黑盒和白盒对抗攻击设置中。我们的发现证实了表格数据也容易受到细微干扰的影响，强调金融技术从业者需要提高对 ML 模型安全性和可信度的认识。此外，通过将基于梯度的攻击方法生成的对抗样本转移到非基于梯度的方法进行实验，也验证了我们的研究结果。我们的结果显示此类攻击仍然有效，突显了开发用于 CCFD 算法的强大防御机制的必要性。

***Index Terms*—对抗攻击，信用卡欺诈检测（CCFD），表格数据，迁移性，金融服务**

I. 介绍

我们正目睹机器学习 (ML) 在金融服务中的广泛采用，这已经彻底改变了欺诈检测、风险评估和反洗钱 (AML) 等领域。迄今为止，这些金融服务主要依赖于传统的 ML 模型，如决策树、逻辑回归和集成学习技术。作为金融安全服务中 ML 的关键应用之一，信用卡欺诈检测 (CCFD) 是一个重要的用例，它能够实现实时识别欺诈交易并增强欺诈预防策略。由于金融机构现在严重依赖这些由 ML 驱动的风险评估、异常检测和欺诈检测系统，确保它们在对抗敌对威胁方面的稳健性已成为一个迫切的问题 [1]。

另一方面，现有的研究表明，在图像数据上训练的 ML 模型存在漏洞，特别是通过向输入数据引入不可察觉的扰动来欺骗 ML 模型。这些攻击通常利用视觉特征的结构化特性来制造对抗样本 [2]。许多研究集中在计算机视觉方面的这些对抗性攻击，但在诸如 CCFD 等金融服务业应用中则较少探索，这些服务一般使用表格数据集。对于在金融应用中常用的表格数据来说，由于其离散、异质的特征分布和领域特定的约束条件 [3]，它带来了独特的挑战。这种研究缺乏带来显著的安全风险，因为金融机构广泛部署了基于 ML 的欺诈检测模型，而没有充分了解它们对对抗性操纵的脆弱性。统计报告显示，在 2020 年仅英国一地，信用卡诈骗就造成了约 5.742 亿英镑的损失 [4]。现有对抗性攻击方法在多大程度上可以应用于用于金融决策的基于表格数据训练的 ML 模型尚不清楚，这给保障基于 ML 的欺诈检测系统留下了关键的安全空白。

为解决这一问题，本文首先考察了在 CCFD 领域中使用最广泛的基于梯度的 ML 模型的安全性。我们采用基于梯度的攻击算法生成对抗样本，并研究它们对欺诈检测性能的影响。为了突出超越传统基于梯度架构的更广泛安全威胁并揭示额外的攻击向量，我们调查了这些对抗样本是否也能误导非基于梯度的 ML 模型。这一调查揭示了 CCFD 模型中的关键漏洞，证明在表格数据上训练的 ML 模型容易受到黑盒和迁移对抗攻击。我们的研究结果强调了迫切需要稳健的防御机制来增强

基于 ML 的欺诈检测系统的安全性。

总之，本文做出了以下关键贡献：

- 我们提出了一个可迁移的对抗攻击的整体框架，特别是针对 CCFD。我们的目标是研究输入样本中的细微扰动如何误导欺诈检测上下文中的机器学习模型的问题。为此，我们评估了一个基于梯度的模型，揭示表格数据可能成为对手易受攻击的目标。（第 III 节）
- 我们使用真实世界的信用卡交易数据集对提出的框架进行了全面评估。此外，我们还评估了对抗样本在未见过的机器学习模型上的有效性，这些模型具有完全不同的模型架构和训练过程。本文中，我们考虑了一个非梯度基础的机器学习模型来评估对抗攻击的可转移性。基于攻击成功率收集的结果表明，我们的工作证明了在白盒设置下生成的对抗样本可以成功转移到结构上不同的模型并保持高攻击效力。（第 IV 节）
- 通过该框架，我们提供了关于 CCFD 模型安全漏洞的经验性见解，揭示了它们对对抗攻击的脆弱性，包括黑盒攻击和转移攻击。我们的研究结果强调了加强金融欺诈检测系统安全性以及开发更强大的防御措施以应对对抗威胁的紧迫需求。（第 V 节）

本文的其余部分结构如下：第 II 节回顾了相关工作。第 III 节介绍了我们提出框架中使用的技术。第 IV 节详细说明了实验设置。第 V 节展示了实验结果，并对其影响进行了深入分析。最后，第 VI 节总结了论文并概述了潜在的未来研究方向。

II. 文献回顾

信用卡欺诈检测（CCFD）是机器学习（ML）在金融安全中的一个重要应用，其中传统的基于规则的系统已被逻辑回归、朴素贝叶斯、KNN、随机森林和支持向量机等监督模型所取代，以提高检测性能 [5]–[7] 。深度神经网络也被采用来进一步提升检测准确性 [8] 。然而，这些模型主要关注分类性能，而它们对对抗威胁的鲁棒性仍处于研究不足的状态 [9] 。对抗攻击——旨在误导模型预测的设计扰动——在基于图像的领域中通过诸如快速梯度符号方法（FGSM）、投影梯度下降（PGD）和 Carlini & Wagner 攻击等方法进行了广泛的研究 [10]–[12] 。然而，它们在表格数据中的适应性，特别是在信用卡欺诈检测中，由于特征的离散性和异质性带来了独特的挑战 [3] 。尽管一些研究探讨了 CCFD 中的对抗风险，但这些工作大多集中在白盒假设下的深度学习上 [13], [14] ，并且渗透测试似乎不足以应对此类威胁 [15] 。与此同时，像逻辑回归和随机森林这样的传统模型—因其简单性和效率仍受青睐用于欺诈检测—在这一领域几乎没有得到关注 [16] 。此外，在金融服务业中将可转移的对抗攻击应用于黑盒环境的可行性尚未得到验证，这是本研究旨在解决的核心研究空白。

III. 提议的框架

本工作的目标是研究金融领域中基于机器学习 (ML) 的信用卡欺诈检测 (CCFD) 模型的安全性。鉴于该领域的相对新颖，现有的关于其安全漏洞的研究仍然有限。基于 ML 算法的 CCFD 通常被表述为一个二元分类问题，在这个问题中，交易会被根据提取出的特征（如交易时间、交易金额、频率和类型）分类为合法或欺诈（例如，信用卡窃取）。ML 模型将使用预定义的决策阈值来进行这些分类。对于对抗性威胁，如果攻击者获得了模型的决策边界或其固有架构和参数的知识，他们可以通过引入沿着梯度方向的小而精心设计的扰动来利用这一信息。这些对抗性扰动可以适当地改变欺诈概率，使一个欺诈交易低于分类阈值，导致误分类为合法交易。

在本节中，我们将首先呈现生成对抗样本的整体框架，随后设计白盒和黑盒对抗攻击。如图 1 所示，我们首先介绍一种对抗攻击方法——快速梯度符号法 (FGSM)，用于为基于梯度的机器学习模型生成对抗样本。目的是评估在 CCFD 背景下对抗攻击是否仍然有效，特别是对于表格数据集。为了进一步研究该领域的潜在攻击向量，我们提取成功绕过第一阶段检测的对抗样本，并将它们用作针对非基于梯度机器学习模型的攻击输入。这一步骤检验了即使目标模型的内部结构和参数完全未知时，可转移攻击是否仍能破坏 CCFD 系统。

通过这一框架，我们旨在揭露机器学习算法及用于欺诈检测的底层表格数据中的漏洞。我们的研究结果强调，在金融行业中开发和部署基于机器学习的欺诈检测模型时，整合安全措施以减轻对抗性威胁的必要性。

A. 机器学习模型在我们的框架中

1) 基于梯度的机器学习： 大多数基于梯度的机器学习模型依赖于梯度下降及其变体来优化模型参数。对于 CCFD，逻辑回归（LR）是一种广泛使用的基于梯

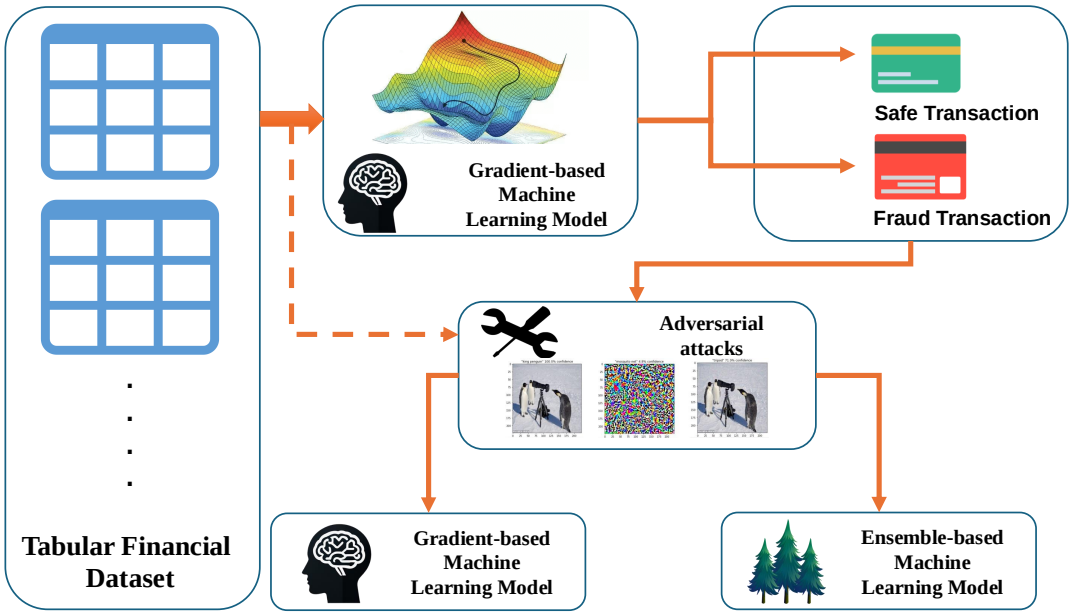

图 1. 基于 FGSM 的信用卡欺诈检测模型对抗攻击的整体框架

度的机器学习算法，它根据输入特征估计事件发生的概率 [17]–[19]。与建模连续输出的线性回归不同，LR 应用逻辑函数将预测映射到 0 和 1 范围内的概率 [20]。逻辑函数定义为：

$$\text{Logit}(\pi) = \frac{1}{1+\exp(-\pi)} \tag{1}$$

其中 π 表示输入特征和模型参数的线性组合。

模型参数使用最大似然估计（MLE）进行优化，以最小化分类误差。训练完成后，LR 预测概率，并通过基于阈值分配标签来进行分类（例如，$\pi \geq 0.5$ 被分类为类别 1，否则为类别 0）[20]。S 形函数刻画了 LR，确保预测值限定在概率范围内。

2) 非梯度基于的机器学习： 一种不同类型的机器学习模型是非基于梯度的模型，与基于梯度的同类相比，它们具有完全不同的架构和训练过程。这些模型通常根据基尼不纯度或熵信息等标准选择特征并进行分类。一个示例模型是随机森林（RF），这是一种用于分类和回归任务的集成学习算法。它构建多个决策树，每个决策树都在数据集和特征空间 [21] 的随机选取子集中训练。在随机森林中通过一种称为自助聚合（bagging）的过程引入了随机性，在这个过程中使用训练数据的不同样本来构建各个树。

对于分类任务，RF 通过所有决策树应用多数投票方案来确定最终预测。在回归任务中，它计算来自所有树的平均预测值。在数据采样和特征选择级别引入随机性可以增强模型泛化能力并减轻过拟合的风险 [22]。

B. 对抗攻击技术在我们的框架中

虽然文献中存在多种对抗攻击技术，本研究关注的是 CCFD 背景下的可转移对抗攻击。为了实现这一目标，我们采用 FGSM 作为主要的攻击技术。FGSM 由 Goodfellow 等人 [10] 引入，展示了 ML 模型对对抗扰动的脆弱性。它通过计算损失函数相对于输入数据的梯度，并在梯度方向上应用一个小扰动来生成对抗样本：

$$\eta = \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{2}$$

其中，ϵ 是一个小扰动因子，θ 表示模型参数，x 是输入，y 是目标标签，$J(\theta, x, y)$ 是损失函数。

扰动的对抗样本计算如下：

$$x^* = x + \eta = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{3}$$

其中 x^* 表示生成的对抗样本。通过利用模型的梯度，FGSM 有效地改变了输入特征以欺骗分类器，同时保持感知变化最小。

IV. 实验

A. 数据集

我们使用了一个来自 Kaggle 的公开信用卡交易数据集 [23]，其中包含 2013 年两天内欧洲持卡人的

284,807 笔交易。该数据集高度不平衡，只有 0.17%（492）被标记为欺诈行为，反映了现实世界的情况。为了保护机密性，特征使用 PCA 进行了匿名处理，形成了变量 V1 – V28，而 'Time'和 'Amount'则得以保留。为了解决类别不均衡问题，我们应用了 SMOTE 技术 [24] 来增加欺诈案例的样本量。数据集通过分层抽样被划分为 80%的训练集和 20%的测试集，以保持类别分布。

B. 评估指标

为了全面评估模型在不平衡的信用卡欺诈检测（CCFD）任务中的性能，我们报告了标准指标包括**准确性**、**精度**和**回忆**。准确性反映了整体预测正确性，而精确率和召回率则捕捉到模型正确识别欺诈案例并最小化假阴性的能力。此外，我们引入了迁移率来评估从一个模型生成的对抗样本误导另一个模型的程度：

$$\text{Transferability Rate} = \frac{\text{Misclassified Samples in Target Model}}{\text{Total Adversarial Samples}}$$

该指标对于评估多分类器欺诈检测场景中的安全风险至关重要。

C. 实验设计

我们采用逻辑回归（LR）作为基线模型，因为它具有较高的可解释性和在金融欺诈检测中的相关性。其依赖基于梯度的优化使其容易受到诸如快速梯度符号方法（FGSM）等对抗攻击的影响。该模型使用 scikit-learn 库 [25] 实现，并作为主要的攻击目标。

为了评估模型的鲁棒性，在拥有完整访问模型梯度权限的白盒设置中，使用对抗鲁棒性工具箱（ART）[26] 对 FGSM 攻击进行了应用。对抗扰动针对测试集中正确分类的欺诈交易，目标是通过最小输入变化将其翻转为非欺诈标签。

为了评估可迁移性，我们测试了为 LR 设计的对抗样本是否可以误导在同一数据集上以 80/20 分层拆分训练的非基于梯度的模型——随机森林（RF）。这种基于转移的黑盒攻击揭示了跨模型漏洞，并突显了异构欺诈检测系统中对抗样本的更广泛风险。

V. 结果与分析

我们展示了对用于信用卡欺诈检测（CCFD）的逻辑回归（LR）模型进行快速梯度符号方法（FGSM）攻击的结果，分析了在不同 ϵ 下召回率下降的情况以及对抗样本向非基于梯度的模型（随机森林）的迁移性。

A. 基线模型性能

为了建立评估对抗攻击影响的参考点，我们首先在 Kaggle CCFD 数据集上训练并评估了一个 LR 模型。基线模型的评估指标如下表 I 所示：

表 I
模型评估指标

度量	值
Accuracy	0.99
Precision	0.17
Recall	0.92

高召回值表明该模型在正常条件下能够有效检测欺诈交易。然而，相对较低的精确度表明该模型产生了大量误报，鉴于欺诈检测中召回率的优先级，这种权衡是可以接受的。

B. FGSM 对逻辑回归的影响

为了评估 LR 模型对对抗攻击的脆弱性，我们使用 FGSM 生成了对抗样本，目标是测试集中正确分类的欺诈交易。将 ϵ 值设置为 2.2，并用生成的对抗样本替换了其良性对应样本。模型在对抗测试集上的性能如下表 II 所示：

表 II
模型评估指标

度量	值
Accuracy	0.99
Precision	0.11
Recall	0.56

与原始召回率 0.92 相比，召回率显著下降，表明近 40%的欺诈交易被错误分类为非欺诈。这种欺诈检测能力的大幅降低突显了对抗性攻击带来的安全风险，强调了在 CCFD 系统中需要有强大的防御机制。

C. epsilon 对模型鲁棒性的影响

为了分析不同对抗扰动幅度对模型性能的影响，我们改变了 ϵ 值并观察其对召回率的影响。结果绘制在图 2 中。

从图 2，我们观察到以下趋势：

- 在 $\epsilon = 0$ 处，召回率为 0.92，表明在没有对抗扰动的情况下正确检测了 92% 的欺诈交易。
- 随着 ϵ 的增加，召回率稳步下降，显示出 FGSM 攻击的有效性不断增强。

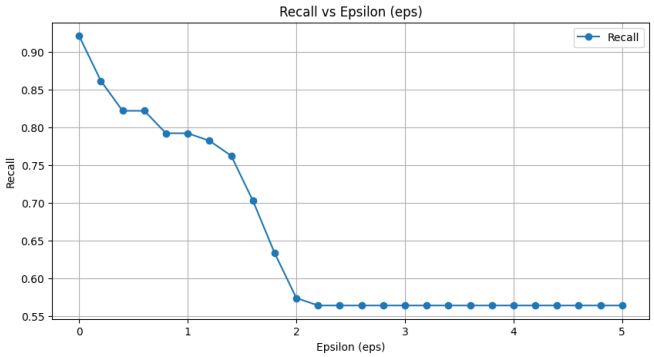

图 2. 对抗扰动增加 (ϵ) 对模型召回率的影响，显示退化趋势及最终稳定。

- 当 $\epsilon = 2.2$ 时，召回率下降到 0.56，这意味着几乎有一半的欺诈交易被误分类。
- 对于 $\epsilon > 2.2$，召回率稳定在 0.56，这表明在此点之后增加扰动幅度不会进一步降低模型性能。

这种饱和效应意味着 FGSM 具有一个最大攻击效果阈值，超过该阈值后进一步的扰动不会引入额外的漏洞。一种可能的解释是随着扰动增加，它们可能会将欺诈样本推得远离其原始分布，使其变得不真实，并且通过异常检测机制可能更容易被发现。此外，过度的扰动可能会使样本以不再影响分类的方式移出模型的决策边界，导致攻击效果递减。

D. 特征扰动分析

为了更好地理解对抗扰动对个体特征的影响，我们分析了一笔成功被扰动的欺诈交易的原始值和对抗值。表 III 提供了扰动详情。

表 III 的关键观察揭示了模型在对抗操控下的行为：

- 经过 PCA 转换的特征（V1 – V28）表现出大约为 ±2.2 的一致扰动，这些扰动足以误导逻辑回归模型。由于这些特征源自主成分分析，即使是很小的变化也能显著改变编码表示。
- 命名特征如时间和数量几乎没有变化，表明对抗性扰动针对的是抽象特征空间而非原始交易细节。
- 这些发现强调了模型对转换输入的敏感性，并突出了在欺诈检测管道中的预处理和特征工程阶段考虑对抗鲁棒性的重要性。

E. 对抗样本的迁移性

为了进一步评估对抗样本的影响，我们测试它们转移到另一个分类模型的能力。具体来说，我们评估使用

表 III
对抗扰动分析对于一个成功被错误分类的欺诈样本。

特征	之前（原始）	之后（对抗性）	扰动
Time	48884.0	48886.2	2.2
V1	-2.13905056	-4.3390503	-2.19999974
V2	1.39436766	-0.8056323	-2.19999996
V3	-0.612034895	-2.8120348	-2.19999991
V4	1.04932706	-1.1506729	-2.19999996
V5	-1.16210191	-3.3621018	-2.19999989
V6	-0.768219363	1.4317807	2.20000006
V7	-1.99723740	0.20276256	2.19999996
V8	0.574996543	2.7749965	2.19999996
V9	-0.980831776	1.2191682	2.19999998
V10	-2.49561925	-0.2956193	2.19999995
V11	2.55558915	0.3555892	-2.19999995
V12	-3.53043627	-1.3304362	2.20000007
V13	-1.01623382	1.1837661	2.19999992
V14	-3.45519658	-1.2551966	2.19999998
V15	-0.056363864	2.1436367	2.20000009
V16	-2.46773703	-0.26773703	2.2
V17	-7.14032597	-4.940326	2.19999997
V18	-1.27128002	0.92871994	2.19999996
V19	-0.00172192354	-2.201722	-2.20000008
V20	0.0254265126	2.2254264	2.19999989
V21	0.696954881	-1.5030451	-2.19999998
V22	0.740003045	-1.4599969	-2.19999995
V23	-0.155115249	-2.3551152	-2.19999995
V24	-0.0506074461	-2.2506075	-2.20000005
V25	0.268368293	-1.9316317	-2.19999999
V26	-0.469432841	1.7305672	2.20000004
V27	-0.405813768	-2.6058137	-2.19999993
V28	-0.152170847	-2.352171	-2.20000015
Amount	19.73	17.53	-2.2

FGSM 攻击在逻辑回归模型上生成的对抗样本是否也能欺骗非基于梯度的模型，如随机森林 (RF)。

1) 基线 RF 模型性能: 一个 RF 分类器在与 LR 模型相同的数据库上进行了训练，遵循了同样的 80/20 的训练-测试分割。RF 在良性测试样本上的基线性能如下表 IV 所示：

表 IV
模型评估指标

度量	值
Accuracy	1.00
Precision	1.00
Recall	0.95

如表 V 所示，在正常条件下，RF 在欺诈检测中几乎达到了完美性能。

表 V
基线随机森林模型在对抗攻击前的混淆矩阵。

Actual	Predicted: Non-fraud	Predicted: Fraud
Non-fraud	56861	5
Fraud	0	96

2) 对抗样本传输性实验: 为了检验 RF 对抗样本的鲁棒性，我们将使用 FGSM 生成的 LR 模型的对抗测试集应用于训练好的 RF 模型。结果如下表 VI 所示：

表 VI
对抗攻击结果

度量	**值**
Successful attacks	34
Failed attacks	2
Transferability success rate	94%

高成功率（94%）表明为基于梯度的模型如 LR 设计的对抗性扰动仍能显著降低非基于梯度的模型如 RF 的性能。这表明对抗性攻击不仅限于依赖梯度的模型，还可以在不同类型的分类器之间传递。

3) 可转移性分析: 几个因素导致对抗样本的高度迁移性：

- 特征空间扰动：尽管随机森林不使用梯度，但它依赖于特征重要性进行分类。FGSM 引入的扰动可能会改变关键特征的决策边界，导致误分类。
- 共享决策边界：LR 和 RF 模型都在同一数据集上进行训练，这意味着它们可能会学到类似的决策边界。这增加了对 LR 有效的对抗扰动也会影响 RF 的可能性。
- 结构集成模型的限制：虽然随机森林受益于引导汇聚和特征随机性，但它仍然容易受到系统性改变输入分布的对抗性扰动的影响，从而导致错误分类。

这些结果突显了在现实世界欺诈检测系统中对抗攻击带来的更广泛的安全风险。即使组织部署多个分类器以提高鲁棒性，可转移性仍可能使所有模型暴露于对抗威胁之下。未来的研究应探索对抗训练技术和其他防御机制来减轻这些风险。

VI. 结论

在本文中，我们研究了信用卡欺诈检测（CCFD）模型对对抗攻击的脆弱性，证明通过快速梯度符号方法（FGSM）生成的对抗样本显著降低了模型性能。特别是逻辑回归模型的召回率从 92%下降到 56%，这些扰动对非基于梯度的随机森林模型表现出 94%的转移率，强调了对抗攻击在金融应用中带来的更广泛的安全风险。

尽管已经提出了多种对抗防御机制——如对抗训练、特征正则化和集成防御 [10] ，但这些方法大多数是为非表格数据集开发的。梯度混淆方法旨在防止攻击者利用模型梯度 [27] ，而输入预处理技术如噪声过滤和数据压缩则试图减轻对抗扰动的影响 [28] 。它们在 CCFD 模型中的适用性尚未得到充分研究。

未来的研究应专注于为表格数据集上的机器学习（ML）模型适应对抗防御方法，评估其在各种金融数据集上的有效性，并探索集成多种缓解技术的混合防御策略。

总之，对抗攻击对基于 ML 的欺诈检测构成了系统性的安全风险，超出了基于梯度的模型范围。开发稳健、特定领域的对抗防御措施对于确保金融欺诈检测系统的可靠性和韧性至关重要。

致谢

首位作者在君士坦丁堡运营有限公司（君士坦丁堡）进行此项工作，作为与悉尼大学合作的一项资助研究项目的一部分。作者们感谢君士坦丁堡对本工作的支持。

参考文献

[1] D. Lunghi, A. Simitsis, O. Caelen, and G. Bontempi, "Adversarial learning in real-world fraud detection: Challenges and perspectives," *arXiv*, vol. 2307.01390, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2307.01390

[2] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, vol. 9, pp. 155 161–155 196, 2021.

[3] F. Cartella, O. Anunciacao, Y. Funabiki, D. Yamaguchi, T. Akishita, and O. Elshocht, "Adversarial attacks for tabular data: Application to fraud detection and imbalanced data," *arXiv preprint arXiv:2101.08030*, 2021.

[4] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, "Random forest for credit card fraud detection," in *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, 2018, pp. 1–6.

[5] A. Ali, S. Abd Razak, S. H. Othman, T. A. E. Eisa, A. Al-Dhaqm, M. Nasser, T. Elhassan, H. Elshafie, and A. Saif, "Financial fraud detection based on machine learning: a systematic literature review," *Applied Sciences*, vol. 12, no. 19, p. 9637, 2022.

[6] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit card fraud detection - machine learning methods," in *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 2019, pp. 1–5.

[7] S. Kumar, V. K. Gunjan, M. D. Ansari, and R. Pathak, "Credit card fraud detection using support vector machine," in *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021*. Springer, 2022, pp. 27–37.

[8] R. Asha and S. K. KR, "Credit card fraud detection using artificial neural network," *Global Transitions Proceedings*, vol. 2, no. 1, pp. 35–41, 2021.

[9] F. V. Jedrzejewski, L. Thode, J. Fischbach, T. Gorschek, D. Mendez, and N. Lavesson, "Adversarial machine learning in industry: A systematic literature review," *Computers & Security*, p. 103988, 2024.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015. [Online]. Available: https://arxiv.org/abs/1412.6572

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019. [Online]. Available: https://arxiv.org/abs/1706.06083

[12] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," 2017. [Online]. Available: https://arxiv.org/abs/1608.04644

[13] A. Agarwal and N. K. Ratha, "Black-box adversarial entry in finance through credit card fraud detection." in *CIKM Workshops*, 2021.

[14] Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and B. Xi, "Adversarial support vector machine learning," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2012, pp. 1059–1067.

[15] D. Lunghi, A. Simitsis, and G. Bontempi, "Assessing adversarial attacks in real-world fraud detection," in *2024 IEEE International Conference on Web Services (ICWS)*. IEEE, 2024, pp. 27–34.

[16] P. Tiwari, S. Mehta, N. Sakhuja, J. Kumar, and A. K. Singh, "Credit card fraud detection using machine learning: a study," *arXiv preprint arXiv:2108.10005*, 2021.

[17] T. Wang and Y. Zhao, "Credit card fraud detection using logistic regression," in *2022 International Conference on Big Data, Information and Computer Network (BDICN)*, 2022, pp. 301–305.

[18] A. Mahajan, V. S. Baghel, and R. Jayaraman, "Credit card fraud detection using logistic regression with imbalanced dataset," in *2023 10th international conference on computing for sustainable global development (iNDIACom)*. IEEE, 2023, pp. 339–342.

[19] M. V. Krishna and J. Praveenchandar, "Comparative analysis of credit card fraud detection using logistic regression with random forest towards an increase in accuracy of prediction," in *2022 International Conference on Edge Computing and Applications (ICECAA)*. IEEE, 2022, pp. 1097–1101.

[20] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[22] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, pp. 31–39, 2017.

[23] M. L. G. ULB, "Credit card fraud detection," n.d. [Online]. Available: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321 – 357, Jun. 2002. [Online]. Available: http://dx.doi.org/10.1613/jair.953

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html

[26] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. M. Molloy, and B. Edwards, "Adversarial robustness toolbox v1.0.0," 2019. [Online]. Available: https://arxiv.org/abs/1807.01069

[27] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint*, vol. arXiv:1705.07204, 2017.

[28] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy, "A study of the effect of jpg compression on adversarial images," *arXiv preprint*, vol. arXiv:1608.00853, 2016.