基于深度学习的随机性检验策略研究

陈东昱[1,2]，陈华[1]，范丽敏[1]，付一方[1,2]，王舰[1,2]
（1. 中国科学院软件研究所可信计算与信息保障实验室，北京 100190；2. 中国科学院大学，北京 100049）

摘 要：为了获得更好的检验效果，对基于深度学习的随机性检验策略进行了研究，包括 2021 年欧密会提出的批均化策略和数据单元大小的选择策略。通过给出基于深度学习方法的随机性统计检验模型，理论推导得到 2 个检验策略的统计量分布和检验势表达，并指出：1) 批均化策略虽然能够提升模型预测准确率，但在统计上容易造成第二类错误概率的增大，反而降低了检验势；2) 一般情况下深度学习模型的数据单元越小，取得的检验势越高。基于以上认识，提出了一种新的比特级深度学习模型用于随机性统计检验。该模型应用于线性同余发生器（LCG）算法，相比之前工作，参数量减少至 $\frac{1}{80}$，取得预测优势所需数据减少了 50%以上；拓展应用于 5～7 轮 Speck 算法获得了明显的预测优势，与 Gohr 模型相比，参数量减少至 $\frac{1}{10}\sim\frac{1}{20}$。

关键词：深度学习；随机性；统计检验；随机数发生器；Speck；线性同余发生器；批均化策略
中图分类号：TN918.1
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2023111

Research on test strategy for randomness based on deep learning

CHEN Dongyu[1,2], CHEN Hua[1], FAN Limin[1], FU Yifang[1,2], WANG Jian[1,2]
1. Trusted Computing and Information Assurance Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
2. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: In order to achieve better test performance, researches on the randomness test strategies based on deep learning were conducted, including the batch average strategy proposed by EUROCRYPT 2021 and the selection strategy for data unit size. By introducing the randomness statistical test model based on deep learning methods, the statistical distribution and test power expression of two test strategies were theoretically derived, and it was pointed out that: (i) the batch average strategy could amplify the prediction accuracy of the model, but it was prone to an increase in the probability of the second type of error in statistics, instead reducing the statistical test power; (ii) the smaller data units of the deep model generally obtained the more powerful statistical tests. Based on the above understanding, a new bit-level deep learning model was proposed for randomness statistical tests, which gained the advantage of prediction with 80 times fewer parameters and 50% samples, compared with the previous work on linear congruent generator (LCG) algorithm, and achieved significant prediction advantages with 10～20 times fewer parameters by extending the model to apply to 5～7 rounds of Speck, compared with the model proposed by Gohr.

Keywords: deep learning, randomness, statistical test, random number generator, Speck, LCG, batch average strategy

0 引言

随机数的质量直接关系到密码系统的安全性。伪随机数发生器（PRNG, pseudo random number generator）一般用某个确定的数学算法产生随机数，例如，C 语言常用线性同余发生器（LCG）产生随机数[1]，MATLAB 内置随机数发生器 MT（Mersenne twister）[2]。随机性统计检验是进行随机数质量评

收稿日期：2023−04−07；**修回日期**：2023−06−12
基金项目：国家重点研发计划基金资助项目（No.2020YFA0309704）
Foundation Item: The National Key Research and Development Program of China (No.2020YFA0309704)

估的主要方式，许多国家和组织给出了相关检验标准用于密码产品的测评，例如美国的 SP800-22[3]、德国的 AIS31[4]、我国的随机数检验规范[5]等；随机性统计检验也常被用于评估一些密码算法（杂凑函数、分组密码）的随机性[6]。

近年来，机器学习（尤其是深度学习算法）已成为工业界和学术界的热点技术，在许多领域都有了较广泛的应用。在密码学领域，深度学习技术已经应用于密码算法编码设计、传统密码攻击、现代密码分析、侧信道攻击、随机性评估、隐私保护、网络安全等方向[7]。本文主要对深度学习技术在随机性评估方面的研究感兴趣，同时注意到现代密码分析中的区分器在统计上实际为一个检验统计量，也可作为检验密码算法随机性统计评估手段，因此关注以下 2 个方面的研究。

在密码分析方面，Gohr[8]基于深度残差网络（ResNet）给出了针对轻量级分组算法 Speck32/64 的 5～7 轮的深度区分器，指出该区分器相对差分分布表的预测优势，并进行了 11 轮的密码攻击。该研究工作引起了深度学习技术应用于现代密码算法分析的研究热潮。Benamira 等[9]分析了该区分器的特征表现，并给出了一种批均化策略（将一个批次内的数据单元预测结果的平均值作为批次的预测结果）提高预测准确度。Bao 等[10]利用中立比特技术成功攻击了 12 轮和 13 轮 Speck32/64。Paterson 等[11]利用大量 RC4 密钥流实例学习得到一个 RC4 的单字节和两字节偏差贝叶斯模型，并基于相同明文的百万次加密后得到的密文，恢复出对应的明文。Mishra 等[12]利用一个简单的深度学习模型成功检测到 RC4 第二字节的非随机性。

在随机性评估方面，Savicky 等[13]提出了一种基于强化学习的随机数检测技术，检测 MATLAB 中一些 PRNG 的随机数之间的依赖关系，如 MT。Fan 等[14]利用人工神经网络（ANN）技术检测，证明 π 的比特表示和 MT 都不是随机的。Fischer[15]提出了一种基于长短期记忆（LSTM）模型的新型 PRNG 检测方法，通过定义一个学习 PRNG 的优化器检测随机数发生器的随机数质量。Truong 等[16]基于循环卷积神经网络（RCNN）提出了一种预测型深度学习分析技术，用以考察确定的经典噪声对量子随机数发生器（QRNG）的随机性和不可预测性的影响程度，其提出的基于 LSTM 的深度学习模型对通过 SP800-22 的 15 项检验（LCG 生成的随机数）的测试序列具有比随机猜测更好的预测效果。Li 等[17]在 LSTM 模型中引入时间模式机制（TPA）对白噪声的非确定性随机数发生器和 LCG 进行检测，验证了其相对于 RCNN 的训练数据量优势。Yang 等[18]提出以神经网络方法近似概率分布函数，进行极小熵估计，指出该方法具有准确性和执行效率的优势。Zhu 等[19]利用基于深度学习的时间变化监测技术，结合对极小熵估计的预测模型，给出了时间变化数据的极小熵估计。

以上工作表明，深度学习方法具有比经典统计方法更好的检验效果的潜力。为了获得更好的随机性检验效果，本文对基于深度学习的随机性统计检验策略进行了研究，包括批均化策略和数据单元大小的选择策略。首先，给出了一种基于深度学习方法的随机性统计检验模型，将模型的预测（或分类）正确率作为统计量进行随机性统计检验，得到具有理论支撑的随机性统计判定准则，并实验证明了基于 Gohr 神经区分器的统计检验方法取得了相对于 SP800-22 随机性统计检验包更好的检验效果。在所提模型基础上，对检验策略进行理论研究，通过推导得到各检验策略的统计性质，给出其检验势表达，即出现第二类错误的概率，得到如下结论。1) 批均化策略虽然能够显著提升预测或分类的正确率，然而通过理论推导和随机实验可以看到，统计上其显著减少了样本量，反而使最终检验势降低；2) 更小的预测单元容易得到更高的检验势。基于此认识，本文提出了一种新的比特级深度学习模型，通过一个简单全连接网络构建输入单元与预测单元之间的函数关系，以此验证其不可预测性和相关性。与 Truong 等[16]和 Li 等[17]所提针对 LCG 算法的深度学习模型相比，本文模型参数量减少至 $\frac{1}{80}$，取得预测优势所需数据减少了 50%以上。拓展应用于 5～7 轮 Speck 算法能够获得明显的预测优势，参数量相比 Gohr 模型[8]减少至 $\frac{1}{10}\sim\frac{1}{20}$。

1 相关工作

本节首先简单介绍了主要的深度学习模型。基于深度学习的密码分析或随机性检验方面的成果非常丰富，但考虑到现实检测需求和复现存疑的问题，本文选择了 2 种公认较好的可用于随机性检验

的深度学习模型，包括 Gohr 神经区分器[8]和 Truong 深度学习检测模型[16]。

1.1　深度学习

深度学习可以使计算机通过简单的概念来构建复杂的概念，即通过简单的线性和非线性函数的层层迭代来表达复杂的数学关系。最常见的有前馈神经网络（FNN）、卷积神经网络（CNN）、循环神经网络（RNN）。FNN 一般适用于全局数学关系或特征的寻找，在输入输出的数学关系挖掘中具有较大潜力；CNN 在模式识别、图像特征提取等方面近年来取得了较好的效果，在实际生活中应用广泛；RNN 每层的输出与输入和之前的计算有关，在处理序列数据中具有较大作用，例如语言处理、翻译等。LSTM 模型和门控循环单元（GRU）都是常用的 RNN，其中 LSTM 可以学习较长时间内的依赖关系，擅长处理序列相关问题。RCNN 是循环神经网络和卷积神经网络的组合，一般利用 CNN 提取特征输入 RNN 进行关联分析。在密码学领域的应用中，CNN 常用来寻找模板特征进行区分，RNN 常用来分析序列依赖关系进行预测，下面介绍的 2 个模型分别使用 CNN 和 RCNN 模型。另外，还有很多相关研究工作使用无监督的自编码器模型进行传统密码算法的攻击、去噪等[20]。

深度学习常用激活函数为修正线性单元 ReLU、Sigmoid 等，常用优化函数有 Adam、RMSProp 等，计算引擎有 Pytorch、Tensorflow、Keras 等。本文实验使用以 Tensorflow 为后端引擎的 Python 深度学习框架 Keras，计算机配置为 Windows 10 Intel Xeon CPU 和两块 NVIDIA RTX3090 GPU。

1.2　Gohr 神经区分器

Gohr[8]提出利用深度学习提升对减轮 Speck32/64 的攻击效果。该工作是第一次用深度神经网络和经典密码分析工具解决相同的问题并进行性能比较，同时是第一篇将神经网络与经典密码分析工具相结合的论文，也第一次给出了基于神经网络的针对对称密码算法的攻击，且攻击效果相比传统最优的结果有所提升。其通过构造多层神经网络得到了一个神经区分器，通过输入密文差分对（通过差分为 0x0040/0000 的明文对进行 5～7 轮加密得到），可以区分 5～7 轮密文对数据与随机数据，并在此基础上可以攻击 11 轮 Speck32/64。

统计上区分器是一个检验统计量，因此 Gohr 构造的神经区分器网络具有改造为随机性统计检验的潜力。Gohr 神经区分器网络模型（后文简称为 Gohr 模型）如图 1 所示，其主体采用了深度残差网络 ResNet，其中，Cov1D 为一维卷积层，Dense 为全连接层，Flatten 为一维展开层，Sigmoid 层为二分类输出函数。

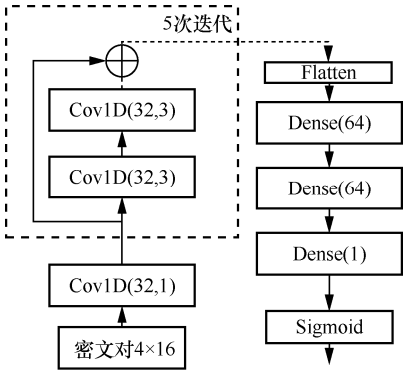


图 1　Gohr 神经区分器网络模型

Gohr 神经区分器主要用于区分密文对和随机数据对，利用神经网络学习固定明文差分 0x0040/0000 下密文对的分布情况，从而对输入数据进行分类判定。通过模型计算可以得到将密文对数据和随机数据分类正确的频率[8]，只要其值远大于 50%，即可作为一个显著的区分器。Benamira 等[9]尝试对该区分器进行解释，并指出将数据分割为小批次，以批次内数据的模型平均得分对其进行分类判定，则可得到更高的模型正确分类概率，从而能获得更有效的评价。另外，也可通过该模型对明文输入差分进行初步筛选，即该模型具有自动化密码分析的潜力，且不需要已知算法的详细知识。这为自动化统计检验提供了较好的思路，同时也需验证其相对于现有统计检验方法是否具有优势。

1.3　Truong 深度学习检测模型

Truong 等[16]提出一种针对 QRNG 的深度学习分析方法，用以考察确定的经典噪声对 QRNG 的随机性和不可预测性的影响程度，即考察以下 2 种环境：1) 无量子源，只有噪声；2) 有量子源和噪声。通过构建一个基于 LSTM 的 RCNN 模型，输入多个相邻 N bit 数据，预测下一个 N bit 输出数据，由此得到模型的预测准确率 P_{ML}，并与随机猜测概率 P_{guess}（通过极小熵可知）对比，若 $P_{\mathrm{ML}} >> P_{\mathrm{guess}}$ 则取得了预测优势。然后，考虑将模型直接应用于 LCG，其基本原理可表示为

$$X_{n+1} = (aX_n + c) \bmod M$$

其中，选择参数 $a=1103\,515\,245$, $c=12\,345$，$M\in\{2^{24},2^{26},2^{28},2^{30}\}$，生成 2.5 亿个 8 bit 数据。输入 10 个 8 bit 数据预测下一个 8 bit 输出数据，当 $M=2^{24},2^{26},2^{28}$ 时能够获得预测优势，即 $P_{\mathrm{ML}}>>0.39$。特别地，当 $M=2^{28}$ 时，所生成的伪随机数可以通过 SP800-22 中所有随机性统计检验项目[16]。由此可知，该模型相对于现有随机性统计检验方法具有明显优势，且具有挖掘复杂数学关系的潜力。Truong 深度学习检测模型如图 2 所示。

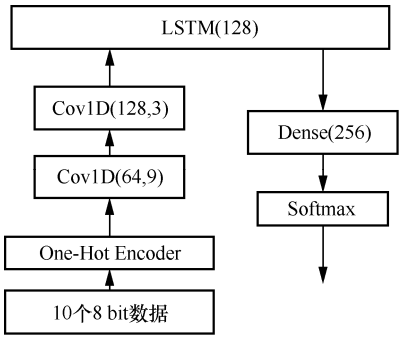


图 2 Truong 深度学习检测模型

图 2 中，独热编码（One-Hot Encoder）层将输入数据变为独热变量，Softmax 层为多分类输出函数。Li 等[17]对 Truong 深度学习检测模型进行了改进，在 LSTM 后加入 TPA，具体模型流程为 One-Hot Encoder + LSTM + TPA + FC。其同样对 LCG 进行检测，取参数 $a=25\,214\,903\,917$, $c=1$，取得了与文献[16]同样的效果。特别地，Li 等[17]对比了 $M=2^{24}$ 时 4 种模型取得预测优势时所需数据量（包括 RNN、FNN、RCNN、TPA），以此说明其模型的效果，验证了该模型相对于文献[16]中 RCNN 模型的训练数据量优势。需要注意的是，当 $M=2^{30}$ 时，2 种模型都没有取得预测优势。

以上 2 种模型采用了不同的检验策略，包括不同的输入数据单元大小和批均化策略等。为了获得更好的检验效果，本文考虑对检验策略进行研究。因此需要构建统计模型，并在此基础上进行理论研究。

2 深度学习统计模型

以上介绍的深度学习模型得到的分类或预测准确率若远大于随机猜测概率，则表示取得了分类或预测概率优势，然而从统计角度看缺少阈值和显著性水平判定等统计指标，即缺少标准统计模型。本节给出一个适用于深度学习模型的统计模型，从而可进行统计推断，形成标准统计检验方法。

2.1 统计模型

首先，进行符号说明。记 Binomial(n,p) 为二项分布，其中，n 为伯努利实验次数，p 为每次成功的概率。记 Normal(μ,σ^2) 为正态分布，其中，μ 为均值，σ^2 为方差。令 E(X) 为随机变量 X 的期望，$D(X)$ 为 X 的方差，$P(A)$ 为事件 A 发生的概率。

本文用深度学习模型预测某个单元，设 X_i 为第 i 个单元的模型预测（或分类）正确性二元变量，即 $X_i=1$ 表示预测正确，$X_i=0$ 表示预测错误。令 $Y=\sum_{i=1}^{n}X_i$ 为 n 个单元在该模型下预测正确的个数，p_{ML} 为模型的预测正确率，即 $p_{\mathrm{ML}}=\dfrac{Y}{n}$，可以将 Y 作为统计量进行分析。

设随机假设下，$P(X_i=1)=p_0$，其中 p_0 可通过极小熵或理论推导得到，则 E$(X_i)=p_0$。由此可得，统计量 Y 服从二项分布，即

$$Y\sim\mathrm{Binomial}(n,p_0) \tag{1}$$

因此，在随机假设下，任何模型的预测准确个数统计量应服从二项分布。可进一步通过近似正态分布来进行概率估计，则有

$$Y\sim\mathrm{Normal}(np_0,np_0(1-p_0)) \tag{2}$$

$$p_{\mathrm{ML}}=\frac{Y}{n}\sim\mathrm{Normal}\left(p_0,\frac{p_0(1-p_0)}{n}\right) \tag{3}$$

由此可以给出显著性水平为 α 的随机性判定策略如下。当 $\dfrac{\alpha}{2}<\Phi\left(\dfrac{p_{\mathrm{ML}}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}\right)<1-\dfrac{\alpha}{2}$ 时，接受随机性假设，否则拒绝。其中，$\Phi(x)$ 为正态分布概率累计函数。则 p_{ML} 的接受（随机性假设）域为

$$\left(\sigma_0\Phi^{-1}\left(\frac{\alpha}{2}\right)+p_0,\sigma_0\Phi^{-1}\left(1-\frac{\alpha}{2}\right)+p_0\right) \tag{4}$$

其中，$\sigma_0=\sqrt{\dfrac{p_0(1-p_0)}{n}}$。例如，使用模型对 1 bit 数据进行预测，取 $n=1\,000$, $\alpha=0.01$，均匀随机假设下 $p_0=\dfrac{1}{2}$，则当 $p_{\mathrm{ML}}\in(0.459\,273,0.540\,727)$ 时，接受该假设，否则拒绝。

2.2 二级检验

进一步地，可以参考现有统计检验方法的二级

检验策略，通过多次实验来增加模型检验力度，即可以设定二级检验：通过率检验和均匀性检验。

1) 通过率检验。给定 N 个子序列，令 α 为深度检验的显著性水平，ξ 为子序列通过检验的比例，即 p_{ML} 在接受域内。理想情况下，若随机性假设成立，ξ 应为二项分布随机变量，其均值为 $\mu=1-\alpha$，标准差为 $\sigma=\sqrt{\dfrac{\alpha(1-\alpha)}{N}}$。如果 ξ 在置信区间 $\mu\pm3\sigma$ 之外，则拒绝待检序列随机的原假设。

2) 均匀性检验。由统计理论可知，若 η 是一个服从累积分布 $F(x)$ 的随机变量，则 $F(\eta)$ 服从[0,1]上的均匀分布。所以理想情况下，p 值 $\Phi\left(\dfrac{(p_{\mathrm{ML}}-p_0)}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}\right)$ 应当服从均匀分布。通过将[0,1]区间划分为 k 个相等的子区间，并计算 $W=\sum_{i=1}^{k}\dfrac{\left(D_i-\dfrac{N}{k}\right)^2}{\dfrac{N}{k}}$，可近似计算得到一个新的 p 值 p_T。其中 D_i 表示落入第 i 个区间的 p 值个数。因为 W 依分布收敛到一个自由度为 $k-1$ 的卡方随机变量，如果 $p_T\leqslant0.0001$，则拒绝随机性假设。

2.3　Gohr 模型统计检验

Truong 深度学习检测模型[16]相对于 SP800-22 统计检验方法的优势已被验证，本文考虑基于 2.1 节的统计模型将 Gohr 神经区分器作为一种随机性统计检验方法，以此验证其相对 SP800-22 统计检验方法的优势。

深度学习模型所需数据集一般分为训练集和检验集，训练集用于训练模型找到非随机因素并进行简单验证反馈，而检验集则用于所得到的模型的检验过程。因此统计检验中的样本量对应检验集。本文根据文献[8]中数据安排，用 10^7 个密文对（其中一半为固定明文差分 0x0040/0000 对应的密文，一半为随机数据对）作为训练数据，用 10^6 个密文对进行检验。对 5～7 轮 Speck32/64 算法进行实验，并取 2^{24} bit（与 $\dfrac{32}{2}\times10^6$ bit 数据量相当）密文对数据进行 SP800-22 随机性统计检验，结果如表 1 所示。

SP800-22 随机性统计检验中，若 p<0.01 则拒绝随机性假设，否则接受随机性假设。Gohr 模型统计检验中，若分类准确率介于区间 (0.498 712, 0.501 288)，即样本量为 10^6、显著性水平为 0.01 的置信区间，则接受随机性假设，否则拒绝随机性假设。如表 1 所示，面对固定明文差分所对应的密文对数据，5 轮和 6 轮数据都无法通过 SP800-22 统计检验，7 轮数据通过了检验，但 5～7 轮数据无法通过 Gohr 模型检验。需要说明的是，本文对相同的数据进行了二级检验，得到了与一级检验一致的结论。由此验证了 Gohr 模型检验相对于 SP800-22 随机性统计检验方法的检验效果更好。这说明深度学习技术在随机性检验方面具有巨大的潜力。

表 1　5～7 轮 Speck32/64 随机性统计检验

检验项目		p		
		5 轮	6 轮	7 轮
SP800-2	频数	0.000 000	0.000 000	0.237 403
	块内频数	0.000 000	0.000 000	0.679 483
	游程	0.000 000	0.000 000	0.856 241
	最长 1 游程	0.000 000	0.806 792	0.465 166
	矩阵秩	0.000 000	0.346 675	0.644 905
	离散傅里叶	0.000 000	0.000 000	0.183 176
	非重叠模板匹配	0.000 000	0.000 000	0.945 836
	重叠模板匹配	0.000 000	0.452 246	0.956 599
	通用统计	0.000 000	0.000 136	0.618 025
	线性复杂度	0.639 372	0.026 867	0.464 382
	重叠子序列	0.000 000	0.042 517	0.954 304
	近似熵	0.000 000	0.000 000	0.981 640
	累加和	0.000 000	0.000 000	0.416 601
	随机游动	0.000 000	0.000 000	0.255 535
	随机游动变量	0.000 000	0.000 000	0.564 618
Gohr 分类准确率		0.929 513	0.788 817	0.612 240

3　检验策略研究

本节依托于 2.1 节统计模型对 Benamira 等[9]所提的批均化策略和数据单元大小选择策略进行统计理论分析。

3.1　批均化策略

Benamira 等[9]指出可以通过对一个小批次内的模型分类结果求平均值并将其作为该批次的分类结果，从而有效提升分类准确率。本文讨论这种策略的统计性质是否有助于提升深度学习模型的统计检验效果。

首先，建立数学模型。设实际模型对单元预测正确率（即分类准确率）为 p_1，共有 m 个批次，每个批次有 n 个单元预测。令 $Y_j=\sum_{i=1}^{n}X_{i,j}$ 表示第 j 个批次的 n 个单元预测正确的个数，其中，$X_{i,j}$ 表示

第 j 个批次第 i 个单元预测是否正确，$X_{i,j}$=1 表示预测正确，$X_{i,j}$=0 表示预测错误。由 2.1 节的统计模型可知

$$Y_j \sim \text{Binomial}(n, p_1) \to \text{Normal}(np_1, np_1(1-p_1))$$

为讨论方便，令

$$I_{Y_j} = \begin{cases} 1, & Y_j \geqslant \dfrac{n}{2} \\ 0, & Y_j < \dfrac{n}{2} \end{cases} \tag{5}$$

表示批次内的平均结果，q_1 表示批均化策略下实际模型对该批次的预测正确率，则

$$q_1 = P(I_{Y_j}=1) = \sum_{i=\left\lfloor \frac{n}{2} \right\rfloor+1}^{n} \binom{n}{i} p_1^i (1-p_1)^{n-i} \sim 1 - \Phi\left(\frac{\frac{n}{2} - np_1}{\sqrt{np_1(1-p_1)}} \right) \tag{6}$$

批均化策略预测正确率与单元预测正确率的关系如图 3 所示。从图 3 可以看出，批均化策略能显著提升预测正确率，且批内单元样本越多，提升越大。

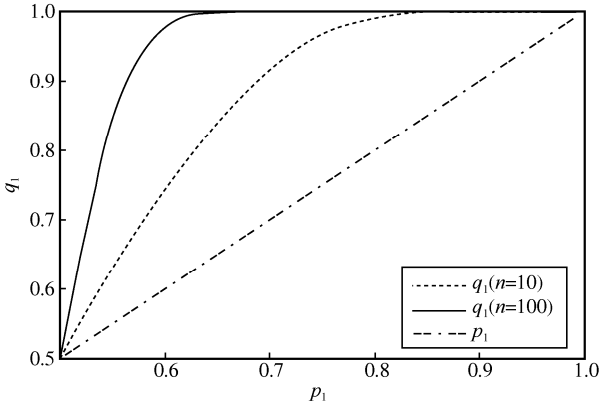


图 3 批均化策略预测正确率与单元预测正确率的关系

下面分析统计检验角度的效果差异。定义 2 个统计量，批均化策略预测正确率统计量和单元预测正确率统计量，分别如式(7)和式(8)所示。

$$Z_1 = \frac{1}{m}\sum_{j=1}^{m} I_{Y_j} \sim \frac{\text{Binomial}(m, q_1)}{m} \to \text{Normal}\left(q_1, \frac{q_1(1-q_1)}{m} \right) \tag{7}$$

$$Z_2 = \frac{1}{m}\sum_{j=1}^{m} \frac{Y_j}{n} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} \frac{X_{i,j}}{n} \to \text{Normal}\left(p_1, \frac{p_1(1-p_1)}{nm} \right) \tag{8}$$

令 p_0 为随机猜测单元正确概率，q_0 为批均化策略下随机猜测批次正确概率，则可以进行统计检验效果的对比，即 2 个检验分别为单元分布（以单元预测结果为随机变量）对比检验和批分布（以批次预测结果为随机变量）对比检验，如式(9)和式(10)所示。

$$\text{Normal}\left(p_0, \frac{p_0(1-p_0)}{nm} \right), \text{Normal}\left(p_1, \frac{p_1(1-p_1)}{nm} \right) \tag{9}$$

$$\text{Normal}\left(q_0, \frac{q_0(1-q_0)}{m} \right), \text{Normal}\left(q_1, \frac{q_1(1-q_1)}{m} \right) \tag{10}$$

为方便讨论，记 $\sigma_i = \dfrac{p_i(1-p_i)}{nm}$，$\sigma_{q_i} = \dfrac{q_i(1-q_i)}{m}$，$i$=0,1。令 $p_0 = 0.5$，$p_1 = 0.52$，共 10 000 个单元预测样本，分为 1 000 个批次，每个批次有 10 个预测单元。则由式(6)可知，批均化策略下批预测正确率为 0.550 368，由此可以对比 2 个检验的概率密度函数（PDF），如图 4 所示。在同一显著性水平 0.01（第一类错误概率 $P(\text{接受}H_1 \mid H_0)$）下，批分布检验出现第二类错误的概率（即 $P(\text{接受}H_0 \mid H_1)$）为 0.193 896（图 4(b)阴影部分面积），远大于单元分布检验出现第二类错误的概率 0.046 968（图 4(a)阴影部分面积）。并且通过实验可知，批次内数据越多，差别越大。

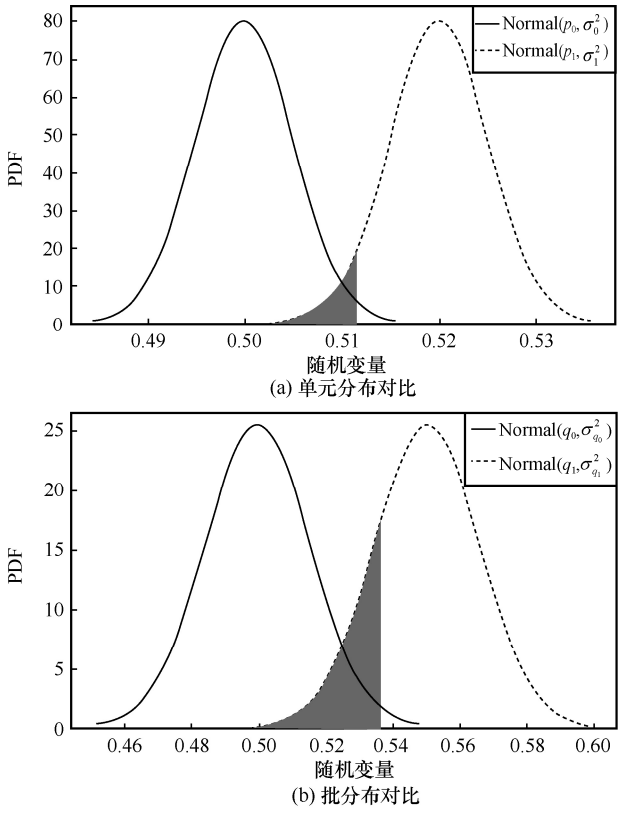


图 4 单元分布对比检验和批分布对比检验

因此可以得出如下结论：批均化策略虽然可以显著提升预测准确率，但减少了检验样本量，导致方差增大，最终使第二类错误出现的概率增加，从统计检验角度，会造成检验势的降低。

3.2　数据单元大小选择策略

前文介绍的基于深度学习的随机性检验可以针对不同的数据对象（字节、N bit 等）来建立模型进行随机性检验。下面讨论在同一检验数据集下，对不同单元大小的深度学习模型统计检验是否有明显检验势差别。

首先，建立数学模型。设总样本量为 n bit，模型预测 N_1 bit 的准确度为 q_1，随机假设情形下为 p_1；模型预测 N_2 bit 的准确度为 q_2，随机假设情形下为 p_2。不妨设 $N_1 < N_2$，则 $p_1 = 2^{-N_1} > p_2 = 2^{-N_2}$。令 $m_1 = \dfrac{n}{N_1}$，$m_2 = \dfrac{n}{N_2}$，则可得如下 2 个检验。

检验 1 为

$$H_0 : N\left(p_1, \frac{p_1(1-p_1)}{m_1}\right), H_1 : N\left(q_1, \frac{q_1(1-q_1)}{m_1}\right) \tag{11}$$

检验 2 为

$$H_0 : N\left(p_2, \frac{p_2(1-p_2)}{m_2}\right), H_1 : N\left(q_2, \frac{q_2(1-q_2)}{m_2}\right) \tag{12}$$

设显著性水平（第一类错误概率）为 α（这里不妨设 $q_1 > p_1,\ q_2 > p_2$，即假设模型具有预测优势），$u_a = \Phi^{-1}(a)$ 为分位数表示，则 2 个检验的第二类错误概率分别为

$$\beta_1 = P\left(\text{接受} H_0 \mid H_1\right) =$$

$$P_{x\sim N\left(q_1, \frac{q_1(1-q_1)}{m_1}\right)}\left(|x - p_1| < u_{1-\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{m_1}}\right) =$$

$$\Phi\left(\frac{p_1 + u_{1-\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{m_1}} - q_1}{\sqrt{\frac{q_1(1-q_1)}{m_1}}}\right) -$$

$$\Phi\left(\frac{p_1 + u_{\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{m_1}} - q_1}{\sqrt{\frac{q_1(1-q_1)}{m_1}}}\right) \tag{13}$$

$$\beta_2 = \Phi\left(\frac{p_2 + u_{1-\frac{\alpha}{2}}\sqrt{\frac{p_2(1-p_2)}{m_2}} - q_2}{\sqrt{\frac{q_2(1-q_2)}{m_2}}}\right) -$$

$$\Phi\left(\frac{p_2 + u_{\frac{\alpha}{2}}\sqrt{\frac{p_2(1-p_2)}{m_2}} - q_2}{\sqrt{\frac{q_2(1-q_2)}{m_2}}}\right) \tag{14}$$

在同一个第一类错误概率 α 下，寻求出现第二类错误概率 β 最小的检验，不出现第二类错误的概率为势，其所对应的检验称为最优势检验。因此，要找到具有更好检验势的参数，就需要分析 $\dfrac{\beta_1}{\beta_2}$ 与 $\dfrac{q_1}{q_2}$ 之间的关系。一般情况下，模型预测单元的长度越大，则样本量越小，对检验效果影响越大，但如果针对大单元的模型能够取得更好的优势，即 q_2 取值更大时，大单元模型的检测能够有更小的第二类错误概率。

本文进行了随机实验。取样本量 $n = 10^6$，单元大小 $N_1 = 8$，$N_2 = 16$，则有 $p_1 = 2^{-8}$，$p_2 = 2^{-16}$，$m_1 = 125\,000$，$m_2 = 62\,500$。设模型取得优势的预测概率 $q_1 = p_1 + 4\sqrt{\dfrac{p_1(1-p_1)}{m_1}} = 0.004\,612$，对 q_2 从 $q_1^{\frac{N_2}{N_1}} \approx 2.127\text{e}^{-5}$（比特独立假设下概率）开始遍历，计算 β_1和β_2。

不同单元大小的检验势（不出现第二类错误的概率）对比如图 5 所示。由于样本量减少，因此 q_2 在比特独立假设下得到的 $q_1^{\frac{N_2}{N_1}}$ 和近似同等概率优势得到的 $p_2 + 4\sqrt{\dfrac{p_2(1-p_2)}{m_2}} \approx 7.775\,83\text{e}^{-5}$，都有 $\beta_1 < \beta_2$，即检验势随预测单元增大而降低；在本文实验中，当 q_2 取值为 $5q_1^{\frac{N_2}{N_1}}$ 或 $p_2 + 6\sqrt{\dfrac{p_2(1-p_2)}{m_2}}$ 附近时，两者的第二类错误概率相近。

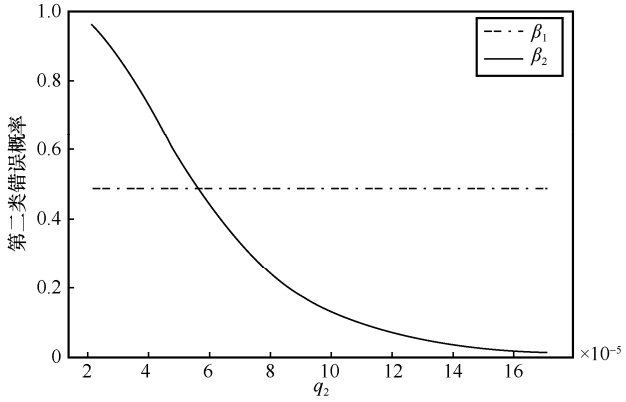


图 5　不同单元大小的检验势对比

从以上讨论可知，在同一数据集下，一般情况下小单元的检验势更佳。因此，同一深度学习模型下，可以尝试从小单元入手寻找非随机因素，更容易获得统计检验上的优势。

4 所提深度随机性检验方法

通过对检验策略的研究，本文认识到从较小的数据单元入手构建深度学习模型更容易获得统计检验上的优势，因此考虑构建比特级数据单元的深度网络用于随机性检验，并验证其效果。

4.1 所提深度检验模型

本节给出一种直观的、可调的深度学习模型，考虑到前文关于检验策略的讨论，可直接针对比特之间的相关性、不可预测性等进行检验。所提深度检验模型如图6所示。

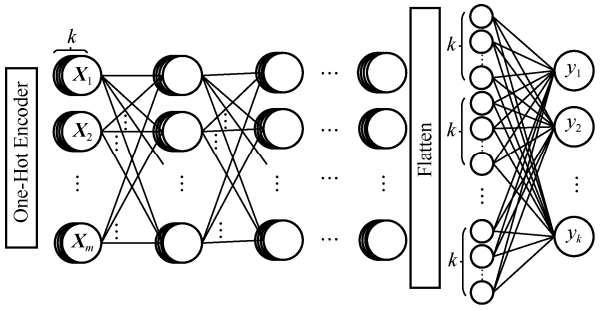


图6 所提深度检验模型

如图6所示，假设一个输入数据单元$\boldsymbol{X}$共有k种可能的值，即模型数据单元大小为lb$k$ bit，将数据单元进行独热向量表示，每个单元用k维向量表示，即$\boldsymbol{X}=(x_1,\cdots,x_k)$。目标建立$m$个输入单元$\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_m$与一个输出单元$\boldsymbol{Y}=(y_1,\cdots,y_k)$之间的数学关系，从而对输出单元的值进行预测。对输入数据特征（即m个k维独热向量）进行l层全连接学习，最后一层的m个单元展开得到km个神经元，与输出单元的k维向量进行全连接计算，对$(y_1,\cdots,y_k)$使用Softmax函数输出得到$\boldsymbol{Y}$预测值结果。例如，将数据单元大小设置为1 bit，则$k=2$，模型用m bit通过多层神经网络来预测某个目标比特。损失函数采用交叉熵CrossEntropy，优化函数可选Adam、RmsProp等。

前文提到了2种可用于随机性检验的深度检验模型，即Gohr神经区分器[8]和Truong深度学习检测模型[16]，其中，Gohr神经区分器主要适用于Speck32/64等分组宽度较小的算法，学习差分模型分布情况；Truong深度学习检测模型可用于迭代型算法检验，主要考察单元的不可预测性。2种模型都使用了CNN搭配LSTM和全连接网络，而本文模型简单来说是一个FNN。密码算法并不像图像或文字那样存在一些明显参数特征（例如图像相邻位置的像素相近等）或长时间的依赖，模型应该建立在对密码算法的数学理解之上，因此最基本的认识便是输入输出各比特之间都存在相关性，本文尝试通过深度学习来得到某些相关性强的数学关系，选择全连接模型，从可能存在的数学关系出发确定输入输出的选择，由此若获得预测概率优势则证明了单元之间的相关性，即本文模型直观解释了相关性强的数学关系，从而做出随机性判定。

下面将本文模型用于LCG算法和Speck算法的随机性检验。

4.2 LCG算法的随机性检验

本节基于4.1节的深度检验模型确定模型参数。单元大小为1 bit，则k取2；输入为32 bit数据，形成一个输入句子，即m取32。每隔8 bit得到一个32 bit句子，中间迭代深度为3，32 bit数据后的下一个比特作为输出。模型具体为Dense(32)+Dense(32)×3+Flatten+Dense(2)，模型整体参数量为4 354个。

Li等[17]对比了LCG算法在$M=2^{24}$时RNN、FNN、RCNN[16]、TPA[17]4种模型取得预测优势所需的数据量，结果表明，4种模型在训练数据量为1.6×10^6 B时，都无法获得预测优势，在数据量达到3.2×10^6 B时FNN和TPA取得了明显预测优势。4种模型都是针对字节单元的预测，最终预测准确率均在随机猜测概率为0.39附近。本文对Truong[16]的RCNN模型进行了验证，同样无法获得任何预测优势，其参数量为336 832。因此，本文选择参数a=1 103 515 245、c=12 345和M=2^{24}，数量为1.6×10^6 B，训练句子数为1 599 996，数据的80%用于训练，20%用于验证，LCG算法的随机性检验训练过程如图7所示。

从图7可以看到，在15个训练周期后，本文模型就取得了明显预测优势，训练数据预测准确度约为0.809 0。注意，训练数据是以10为随机种子生成的，本文以139为随机种子生成10^6 B检验数据，分为5组进行检验，每组399 992个句子，即一次检验样本量，预测准确率分别为65.52%、65.41%、65.45%、65.51%、65.49%。通过2.1节的检验模型可知，显著性水平为0.01的接受域为$(0.497\,964,0.502\,036)$，因此5次检验皆拒绝其随机

性假设。所以本节实验中，本文所提比特深度检验模型与文献[16-17]模型对比，在训练数据量为 1.6×10^6 B 时便取得了预测优势，数据量要求减少了 50%，同时参数规模也从 336 832 个参数减少为 4 354 个参数，减少至 $\frac{1}{80}$。

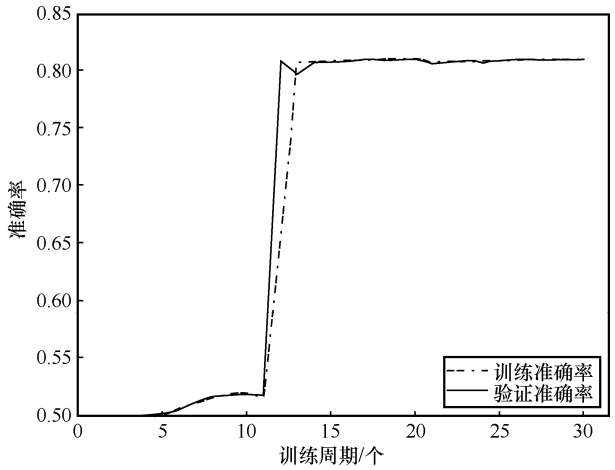
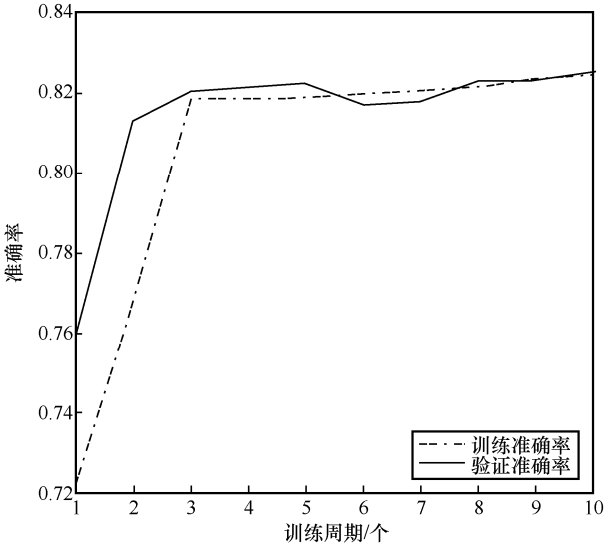


图 7　LCG 算法的随机性检验训练过程

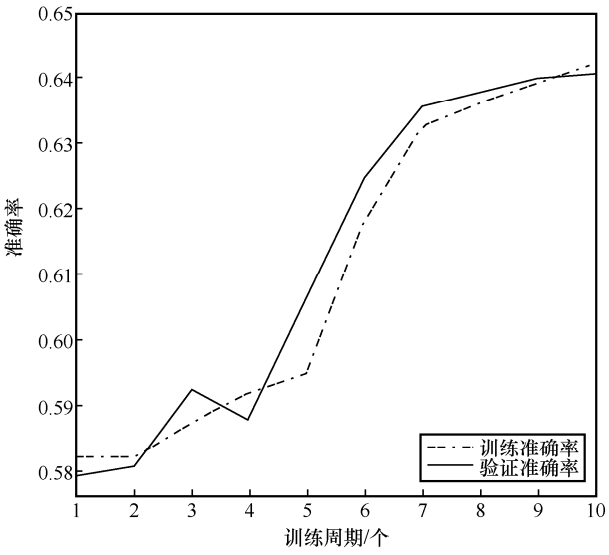
4.3　Speck 算法的随机性检验

本节将本文模型拓展用于 Speck32/64 的随机性检验。具体模型架构与 LCG 算法的随机性检验一致，只是输入变为 31 bit，每隔 32 bit 得到一个句子，输出为 1 bit。与 Gohr 神经区分器[8]一样，取固定明文差分 0x0040/0000 下对应的 5～7 轮加密后的密文对，这里直接将密文对差分作为检验对象，考察 32 bit 密文对差分中的某个比特与其他比特之间的相关性，分别作为模型输出和输入，5～7 轮 Speck 算法的随机性检验训练过程图 8 所示。

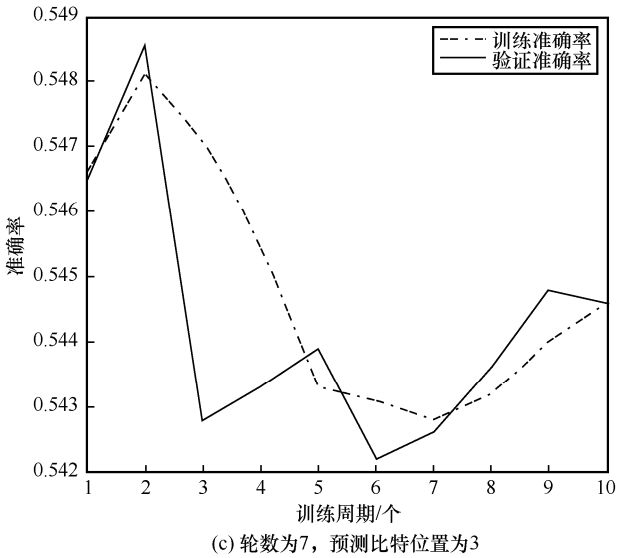
5～7 轮模型训练数据量皆为 10^7 个密文差分对，继续随机生成 10^6 个密文差分对进行随机性检验，得到 5～7 轮的预测准确率分别为 80.49%、64.15%、54.38%。样本量为 10^6，显著性水平为 0.01 的统计模型接受域为 $(0.498\,712, 0.501\,288)$，因此本文模型仅用 10 个周期、4 354 个模型参数便取得了与 Gohr 神经区分器[8]相当的预测概率优势。与 Gohr 神经区分器[8]参数量 102 497 个（5 轮和 6 轮模型）、44 321 个（7 轮模型）相比，本文模型参数量减少至 $\frac{1}{10}\sim\frac{1}{20}$。另外，通过 Gohr 神经区分器[8]无法获知做出随机性判定的原因，Benamira 等[9]尝试对该模型所学到的知识进行解释，相比之下，本文模型可以直观展示输出比特与输入比特之间的相关性。



(a) 轮数为5，预测比特位置为3



(b) 轮数为6，预测比特位置为10



(c) 轮数为7，预测比特位置为3

图 8　5～7 轮 Speck 算法的随机性检验训练过程

本节用本文所提的比特级深度检验模型对 2 种不同密码算法进行了随机性检验，2 种具体模型只是在输入输出结构和部分参数上有所不同。本文模型直接针对输入输出之间的数学关系，可直观解释单元间存在的相关性。本文依据较小数据单元更易获得统计检验优势的认识，使用比特数据单元构建深度检验模型。可以看到，一方面降低了获得统计检验势的数据量要求，另一方面也减少了参数量，减少了模型资源占用。这样的小模型可以更快地进行多方面的学习，基于对密码算法的数学关系认识进行参数调整，得到更具针对性的随机性检验。

5 结束语

本文主要关注基于深度学习方法的随机性检验，并对其 2 个方面检验策略进行了理论研究，包括批均化策略和数据单元大小选择策略。为此，本文首先给出了深度学习统计模型，可对深度学习网络预测模型进行统计表达，并给出其统计性质和随机性判定准则。将 Gohr 神经区分器应用此模型作为一种随机性检验方法，验证了其相对于现有随机性统计检验方法的优势。然后，基于该统计模型，对 2 个方面的检验策略的统计量进行了理论推导和随机实验，并得到结论：批均化预测策略虽然能够放大预测优势，但从统计角度上看会使样本量减少、检验分布方差增大，反而使统计检验势降低，即第二类错误概率增大；深度学习模型输入输出的单元大小，即模型的对象，也对检验势有影响，一般情况下小单元能取得更高的检验势。基于此认识，本文给出了一种比特级深度检验方法，并采用 1 bit 作为模型单元，应用于 LCG 算法和 Speck32/64 算法，与文献[8,16-17]的工作相比，本文模型参数规模小，获得预测优势所需数据量更少，可扩展性强。与文献[8,16]的工作对比可以看到，基于深度学习的随机性检验方法能够取得比现有统计检验方法更强的检验力度，其根本在于非随机复杂数学关系的自动化学习能够获得基于数据的个性化非随机因素。本文模型可直观解释单元间的相关性，但模型依然较初级，一方面参数容易陷于局部最优，另一方面现阶段只是多个单元对一个单元进行预测，多对多的相关性潜力尚未挖掘。因此，如何从随机性统计检验角度设计更具潜力的深度学习模型值得继续研究。

参考文献：

[1] BOYAR J. Inferring sequences produced by a linear congruential generator missing low-order bits[J]. Journal of Cryptology, 1989, 1(3): 177-184.

[2] MATSUMOTO M, NISHIMURA T. Mersenne twister: a 623- dimensionally equidistributed uniform pseudo-random number generator[J]. ACM Transactions on Modeling and Computer Simulation, 1998, 8(1): 3-30.

[3] RUKHIN A, SOTO J, NECHVATAL J. A statistical test suite for random and pseudorandom number generators for cryptographic applications[R]. 2001.

[4] KILLMANN W, SCHINDLER W. Functionality classes and evaluation methodology for true (physical) random number generators: AIS 31[S]. (2001-09-25).

[5] 国家密码管理局. 随机性检测规范: GM/T 0005—2021[S]. 北京: 中国标准出版社, 2021.
State Cryptography Administration. Randomness testing specifications: GM/T 0005—2021[S]. Beijing: Standards Press of China, 2021.

[6] SULAK F. Statistical analysis of block ciphers and hash functions[D]. Ankara: Middle East Technical University, 2011.

[7] ALANI M M. Applications of machine learning in cryptography: a survey[C]//Proceedings of the 3rd International Conference on Cryptography, Security and Privacy. New York: ACM Press, 2019: 23-27.

[8] GOHR A. Improving attacks on round-reduced speck32/64 using deep learning[C]//Proceedings of Annual International Cryptology Conference. Berlin: Springer, 2019: 150-179.

[9] BENAMIRA A, GERAULT D, PEYRIN T, et al. A deeper look at machine learning-based cryptanalysis[C]//Proceedings of Annual International Conference on the Theory and Applications of Cryptographic Techniques. Berlin: Springer, 2021: 805-835.

[10] BAO Z Z, GUO J, LIU M C, et al. Enhancing differential-neural cryptanalysis[C]//Proceedings of International Conference on the Theory and Application of Cryptology and Information Security. Berlin: Springer, 2022: 318-347.

[11] PATERSON K, POETTERING B, SCHULDT J C N. Big bias hunting in Amazonia: large-scale computation and exploitation of RC4 biases[C]// Proceedings of International Conference on the Theory and Application of Cryptology and Information Security. Berlin: Springer, 2014: 398-419.

[12] MISHRA G, GUPTA I, MURTHY S V S S N V G K, et al. Deep learning based cryptanalysis of stream ciphers[J]. Defence Science Journal, 2021, 71(4): 499-506.

[13] SAVICKY P, ROBNIK-ŠIKONJA M. Learning random numbers: a MATLAB anomaly[J]. Applied Artificial Intelligence, 2008, 22(3): 254-265.

[14] FAN F L, WANG G. Learning from pseudo-randomness with an artificial neural network-does God play pseudo-dice?[J]. IEEE Access, 2018, 6: 22987-22992.

[15] FISCHER T. Testing cryptographically secure pseudo random number generators with artificial neural networks[C]//Proceedings of 2018 17th IEEE International Conference on Trust, Security and Privacy In Computing and Communications/ 12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE). Piscataway: IEEE Press, 2018: 1214-1223.

[16] TRUONG N D, HAW J Y, ASSAD S M, et al. Machine learning cryptanalysis of a quantum random number generator[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(2): 403-414.

[17] LI C, ZHANG J G, SANG L X, et al. Deep learning-based security verification for a random number generator using white chaos[J]. Entropy, 2020, 22(10): 1134.

[18] YANG J, ZHU S Y, CHEN T Y, et al. Neural network based min-entropy estimation for random number generators[C]//Proceedings of International Conference on Security and Privacy in Communication Systems. Berlin: Springer, 2018: 231-250.

[19] ZHU S Y, MA Y, LI X S, et al. On the analysis and improvement of min-entropy estimation on time-varying data[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1696-1708.

[20] AHMADZADEH E, KIM H, JEONG O, et al. A novel dynamic attack on classical ciphers using an attention-based LSTM encoder-decoder model[J]. IEEE Access, 2021, 9: 60960-60970.

[作者简介]

陈东昱（1989− ），男，山东淄博人，中国科学院软件研究所博士生，主要研究方向为随机数发生器设计与随机性统计检验。

陈华（1976− ），女，山东日照人，博士，中国科学院软件研究所正高级工程师、博士生导师，主要研究方向为侧信道分析与防护、密码检测。

范丽敏（1978− ），女，内蒙古赤峰人，博士，中国科学院软件研究所高级工程师，主要研究方向为随机性检验、密码检测及侧信道分析与防护。

付一方（1997− ），男，辽宁凌源人，中国科学院软件研究所硕士生，主要研究方向为随机数发生器设计与随机性统计检验。

王舰（1998− ），男，山东临沂人，中国科学院软件研究所博士生，主要研究方向为侧信道分析与防护。