IEICE TRANSACTIONS

on Information and Systems

DOI:10.1587/transinf.2024EDP7011

Publicized:2024/04/12

This advance publication article will be replaced by the finalized version after proofreading.

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Investigating and Enhancing the Neural Distinguisher for Differential Cryptanalysis

Gao WANG[†], **Gaoli WANG**[†,††a)], ***and*** **Siwei SUN**[†††], ***Nonmembers***

SUMMARY

At Crypto 2019, Gohr first adopted the neural distinguisher for differential cryptanalysis, and since then, this work received increasing attention. However, most of the existing work focuses on improving and applying the neural distinguisher, the studies delving into the intrinsic principles of neural distinguishers are finite. At Eurocrypt 2021, Benamira et al. conducted a study on Gohr's neural distinguisher. But for the neural distinguishers proposed later, such as the r-round neural distinguishers trained with k ciphertext pairs or ciphertext differences, denoted as $ND^{cp}_{k\_r}$ (Gohr's neural distinguisher is the special $ND^{cp}_{k\_r}$ with $k = 1$) and $ND^{cd}_{k\_r}$, such research is lacking. In this work, we devote ourselves to study the intrinsic principles and relationship between $ND^{cd}_{k\_r}$ and $ND^{cp}_{k\_r}$.

Firstly, we explore the working principle of $ND^{cd}_{1\_r}$ through a series of experiments and find that it strongly relies on the probability distribution of ciphertext differences. Its operational mechanism bears a strong resemblance to that of $ND^{cp}_{1\_r}$ given by Benamira et al.. Therefore, we further compare them from the perspective of differential cryptanalysis and sample features, demonstrating the superior performance of $ND^{cp}_{1\_r}$ can be attributed to the relationships between certain ciphertext bits, especially the significant bits. We then extend our investigation to $ND^{cp}_{k\_r}$, and show that its ability to recognize samples heavily relies on the average differential probability of k ciphertext pairs and some relationships in the ciphertext itself, but the reliance between k ciphertext pairs is very weak.

Finally, in light of the findings of our research, we introduce a strategy to enhance the accuracy of the neural distinguisher by using a fixed difference to generate the negative samples instead of the random one. Through the implementation of this approach, we manage to improve the accuracy of the neural distinguishers by approximately 2% to 8% for 7-round Speck32/64 and 9-round Simon32/64.

key words: *Differential Cryptanalysis, Neural Distinguisher, Deep Learning, Interpretability, Block Ciphers*

1. Introduction

Lightweight block ciphers, serving as the fundamental components of cryptographic systems, play a pivotal role in safeguarding data confidentiality within resource-constrained devices. Its security has a direct bearing on the safety of our data, underscoring the importance of conducting comprehensive evaluations of the cipher's robustness and reliability. As the field of Machine Learning continues to advance and find broader applications, an expanding cohort of cryptographers has begun to investigate its utility in a range of cryptanalysis techniques. These methods include differential cryptanalysis [1], linear cryptanalysis [2], integral cryptanalysis [3], and Side-Channel cryptanalysis [4], and others. In this paper, we focus on machine learning-based differential cryptanalysis.

At Crypto 2019, Gohr [1] first proposed the concept of neural differential distinguishers and successfully applied it to the reduced-round Speck32/64. Compared to the r-round pure differential distinguishers D_r, the neural differential distinguishers can achieve higher accuracy. However, the reasons behind the superior performance of the neural distinguisher remain unclear. To address this knowledge gap, Benamira et al. [5] explored its underlying principles from both cryptanalysis and machine learning perspectives at Eurocrypt 2021, as detailed in Section 2.5.

In machine learning-based differential cryptanalysis, the success rate of key recovery, as well as the data complexity, is related to the distinguisher's accuracy. Therefore, improving accuracy of the neural distinguisher is an important task. A common strategy to improve the accuracy of neural networks across various machine learning tasks is to provide more features, such as [6,7]. The optimization work in [5] also aims to provide additional features to the neural network. Inspired by this, Chen et al. [8] used k ciphertext pairs to develop a superior r-round neural distinguisher, denoted as $ND^{cp}_{k\_r}$. In fact, Gohr's neural distinguisher is a special case of Chen's neural distinguisher with $k = 1$, i.e., $ND^{cp}_{1\_r}$. In contrast to using multiple ciphertext pairs, Hou et al. [9] employed k ciphertext differences to construct the r-round neural distinguisher $ND^{cd}_{k\_r}$. Additionally, some studies, such as [10, 11], attempted to incorporate information from previous rounds to enhance the accuracy of neural distinguishers. However, none of these works investigated the factors for enhancing the accuracy of the neural distinguisher.

Moreover, as outlined in [5], it is unfair to directly compare D_r and $ND^{cp}_{1\_r}$ due to their different training methods, where $ND^{cp}_{1\_r}$ is trained with the ciphertext pairs, while D_r only relies on their differential probability. Although Benamira et al. constructed $ND^{cd}_{1\_r}$ with ciphertext differences and compared its accuracy with $ND^{cp}_{1\_r}$, they did not conduct a more in-depth investigation, such as the working mechanism of $ND^{cd}_{1\_r}$, the relationships between D_r and $ND^{cd}_{1\_r}$ or between $ND^{cd}_{1\_r}$ and $ND^{cp}_{1\_r}$. The research on the underlying

[†]The author is with the Shanghai Key Laboratory of Trustworthy Computing, Software Engineering Institute, East China Normal University, Shanghai, China

[††]The author is with the State Key Laboratory of Cryptology, Bengjing, China

[†††]The author is with the School of Cyber Security,University of Chinese Academy of Sciences,Beijing, China

a) E-mail: glwang@sei.ecnu.edu.cn

principles of $ND^{cd}_{k\_r}$ and $ND^{cp}_{k\_r}$ is also notably absent.

Our Contributions. In this paper, we aim to analyze the principles and interrelationships of the neural distinguishers $ND^{cd}_{k\_r}$ and $ND^{cp}_{k\_r}$ and explore the methods to enhance the accuracy of the neural distinguisher. We begin our exploration by investigating the operational mechanisms of $ND^{cd}_{1\_r}$ in accordance with the methodologies outlined in [5]. The findings are summarized as follows:

- Choosing the input difference of the optimal r-round differential path can not improve the accuracy of $ND^{cd}_{1\_r}$, similar to its impact on $ND^{cp}_{1\_r}$.
- $ND^{cd}_{1\_r}$ is proficient in identifying the (truncated) differences that exhibit a high bit bias or differential probability within the penultimate or antepenultimate rounds, like $ND^{cp}_{1\_r}$.
- $ND^{cd}_{1\_r}$ has a threshold interval τ, and it can effectively recognize the differences with a probability exceeding the upper limit of τ, but struggles to discern differences with a probability below the lower limit of τ. For the differences whose probabilities fall within τ, $ND^{cd}_{1\_r}$ can only recognize a part of them. This characteristic explains why $ND^{cd}_{1\_r}$ can recognize (truncated) differences that have a high bit bias or probability in the last two rounds.

Subsequently, we delve deeper into the correlation between $ND^{cd}_{1\_r}$ and $ND^{cp}_{1\_r}$. We uncover that $ND^{cp}_{1\_r}$ also has a τ, and its τ is very similar, if not identical, to that of $ND^{cp}_{1\_r}$. Moreover, by confusing the bit relations in the ciphertext pairs, we demonstrate that the $ND^{cp}_{1\_r}$ has the capacity to discern certain relationships among the ciphertext bits, especially the significant bits. This ability is the key to $ND^{cp}_{1\_r}$ achieving superior accuracy over $ND^{cd}_{1\_r}$.

After that, we further broaden these experiments to $ND^{cp}_{k\_r}$ and demonstrate that the average differential probability of k ciphertext pairs plays a crucial role in enabling $ND^{cp}_{k\_r}$ to accurately identify the samples. In contrast, the correlation between ciphertext pairs is very tenuous. Even when a ciphertext pair within a sample is intentionally confused, the $ND^{cp}_{k\_r}$ can still distinguish the sample with the same accuracy as $ND^{cp}_{(k-1)\_r}$. Additionally, the significant bits and importance rankings of the ciphertext bits for both $ND^{cp}_{k\_r}$ and $ND^{cp}_{1\_r}$ are the same.

Finally, according to our research, we propose a scheme to enhance the accuracy of neural distinguisher by employing a fixed difference to generate the negative samples instead of the random one. This strategy efficiently improves the accuracy of the neural distinguisher for 7-round Speck32/64 and 9-round Simon32/64, as shown in Table 6, which clearly demonstrates the effectiveness of our new method.

Outline. The structure of this paper unfolds as follows: Section 2 provides essential background knowledge, including the block cipher Speck32/64 and Simon32/64, the pure differential distinguisher D_r, the neural differential distinguishers $ND^{cp}_{k\_r}$ and $ND^{cp}_{k\_r}$, and a brief description of the works in [5]. In Section 3, we delve into the intrinsic principles of $ND^{cd}_{1\_r}$. Following this, Section 4 undertakes a comparative analysis between $ND^{cd}_{1\_r}$ and $ND^{cp}_{1\_r}$ in terms of cryptanalysis, feature confusion and feature importance. After that, Section 5 further studies the working mechanism of $ND^{cp}_{k\_r}$, proposing a strategy to enhance the accuracy of the $ND^{cp}_{k\_r}$ and applying it to 7-round Speck32/64 and 9-round Simon32/64. Finally, our researches are summarized in Section 6.

2. Preliminary

2.1 Notations

The notations used in this paper are given in Table 1.

Table 1 Notations.

Notation	Description
α	Plaintext difference.
β	Ciphertext difference.
$\oplus$	Bit-wise XOR .
$\boxplus$	Modular addition 2^n
D_r	The r-round pure differential distinguisher.
$ND^{cp}_{k\_r}$	The r-round neural differential distinguisher constructed with the k ciphertext pairs in [8].
$ND^{cd}_{k\_r}$	The r-round neural differential distinguisher constructed with the k ciphertext differences in [9].
$ND^{cp'}_{k\_r}$	The r-round neural differential distinguisher constructed with the k ciphertext pairs in Section 5.4.
Acc	The accuracy of the differential distinguisher.
TPR	The true positive rate of the differential distinguisher.
TNR	The true negative rate of the differential distinguisher.

2.2 The Speck32/64 and Simon32/64 Cipher

Simon and Speck [12] are the lightweight block ciphers proposed by the National Security Agency (NSA) in 2013. Both of them have 10 variants, we focus on the variant with a 32-bit block size and a 64-bit key size in this paper, known as Speck32/64 and Simon32/64. The encryption process of Speck32/64 is executed through 22 rounds, whereas Simon32/64 implements it across 32 rounds. The round function for Speck32/64 involves modular addition 2^n $\boxplus$, left and right circular shift ($\lll$ and $\ggg$), bitwise XOR $\oplus$, whereas it is bitwise XOR $\oplus$, bitwise AND $\wedge$ and left circular shift $\lll$ for Simon32/64. Their encryption processes are given in Algorithm 1 and Algorithm 2, where K^{i-1} and (L^i, R^i) represent the round key and intermediate state for the rounds i.

2.3 Pure Differential Distinguisher

In differential cryptanalysis, cryptographers aim to distinguish a block cipher from the random permutation function by meticulously studying the propagation properties of plaintext difference. For the traditional differential cryptanalysis, the cryptographers focus on searching for the differential path

Algorithm 1 Encryption Algorithm of Speck32/64

Input: $(L^0, R^0) \in \{0,1\}^{32}$
Output: $(L^{22}, R^{22}) \in \{0,1\}^{32}$
for all $i \leftarrow 0, 22$ **do**
  $L^i \leftarrow ((L^{i-1} \ggg 7) \boxplus R^{i-1}) \oplus K^{i-1}$
  $R^i = (R^{i-1} \lll 2) \oplus L^i$;
end for
return (L^{22}, R^{22})

Algorithm 2 Encryption Algorithm of Simon32/64

Input: $(L^0, R^0) \in \{0,1\}^{32}$
Output: $(L^{32}, R^{32}) \in \{0,1\}^{32}$
for all $i \leftarrow 0, 32$ **do**
  $L^i \leftarrow ((L^{i-1} \lll 1) \wedge (L^{i-1} \lll 8) \oplus R^{i-1} \oplus (L^{i-1} \lll 2)) \oplus K^{i-1}$
  $R^i = L^{i-1}$
end for
return (L^{32}, R^{32})

$(\alpha \rightarrow \beta)$ with the best differential probability $DP(\alpha \rightarrow \beta)$, such as [13–15]. For a block cipher E: $F_2^n \rightarrow F_2^n$,

$$DP(\alpha \rightarrow \beta) = \frac{\#\{x | E(x \oplus \alpha) \oplus E(x) = \beta\}}{2^n}, \tag{1}$$

where $\#\{S\}$ represents the number of elements in the set S. The block cipher follows the differential transformation $\alpha \rightarrow \beta$ with the probability $DP(\alpha \rightarrow \beta)$, while it is 2^{-n} for the random permutation function. This property allows the attacker to construct a *differential distinguisher* based on the probability of differential transformation $(\alpha \rightarrow \beta)$.

As opposed to only focusing on a single differential path, Blondeau et al. [16] introduced a method called *multiple differential cryptanalysis*, which considers each differential path in a differential collection Δ, i.e. $\alpha_i \rightarrow \beta_j$, where $\alpha_i \in \Delta$ and $\beta_j \in \Delta$. The *pure differential distinguisher* D_r is a type of the multiple differential cryptanalysis that takes care of each ciphertext differences $\beta_j \in F_2^n$ of a plaintext difference α. For a given differential pair (α, β_j), the D_r recognizes it according to the Equation (2) provided below.

$$\begin{cases} \textit{Encryption cipher } E, \textit{ if } DP(\alpha \rightarrow \beta_j) > 2^{-n} \\ \textit{Random permutation function, otherwise} \end{cases} \tag{2}$$

2.4 Two Different Types of Neural Distinguishers

At Crypto 2019, Gohr [1] first introduced the neural distinguisher to recognize ciphertext pairs from random data based on the fundamental principles of differential cryptanalysis. This breakthrough marked a substantial advancement in the application of deep learning to differential cryptanalysis. Gohr's neural distinguisher employs a single ciphertext pair per sample, each derived by encrypting a plaintext pair that adheres to a fixed difference α. Subsequently, to refine the features available to the neural network, Chen et al. [8] constructed r-round neural distinguishers $ND^{cp}_{k\_r}$ using k ciphertext pairs, as opposed to a single pair. It's worth noting that Gohr's neural distinguisher can be seen as a specific instance of Chen's neural distinguisher with $k = 1$, i.e., $ND^{cp}_{1\_r}$.

In addition to using ciphertext pairs, certain cryptanalysts have also adopted the k ciphertext difference to construct the neural distinguisher $ND^{cd}_{k\_r}$, such as [9, 17]. The only difference between $ND^{cp}_{k\_r}$ and $ND^{cd}_{k\_r}$ is their data formats. Hence, we introduce them together here. Their construction process consists of two phases: the sample generation phase and the training phase. The sample generation phase involves the creation of data samples for $ND^{cp}_{k\_r}$ and $ND^{cd}_{k\_r}$, while the training phase focuses on training the neural distinguisher using these generated samples. The detailed process is as follows:

- **Sample Generation.**
  1. Generate $\frac{N}{k}$ plaintext sets. Half of them consist of k plaintext pairs with difference α, while the other half exhibit randomized plaintext differences. These two types of plaintext set are respectively labeled as 1 and 0.
  2. Get ciphertext sets by encrypting plaintext sets using the encryption cipher E.
  3. For $ND^{cp}_{k\_r}$, its sample consists of the ciphertext sets and their labels.
  4. For $ND^{cd}_{k\_r}$, its sample consists of the set of ciphertext differences and their labels.
- **Training.**
  1. Generate 10^7 training data and 10^6 validation data, respectively.
  2. Feed the training data and validation data into the neural network to train the neural distinguishers.
  3. If the validation accuracy of the $ND^{cp}_{k\_r}$ or $ND^{cd}_{k\_r}$ is greater than 0.5, a available neural distinguisher is acquired, otherwise, it is not.

Neural network architecture. In this paper, our goal is not to explore the design of neural network for neural distinguishers. Therefore, we employ the well-validated residual network architecture as described in [8]. The input layer of $ND^{cp}_{k\_r}$ is the k ciphertext pairs, forming a matrix with dimensions of $k \times 2n$. In contrast, for $ND^{cd}_{k\_r}$, which uses k ciphertext differences as a sample, the input is a matrix with dimensions of $k \times n$. To prevent the risk of overfitting, we set the epoch number to 30 and the model depth to 1. All other parameters remain consistent with those detailed in [8].

2.5 Overview of Benamira's Work

In order to figure out the inner workings of Gohr's neural distinguisher, $ND^{cp}_{1\_r}$, a comprehensive exploration was conducted by Benamira et al. [5] at Eurocrypt 2021, integrating perspectives from both cryptanalysis and machine learning. From a cryptanalysis perspective, they conducted two distinct explorations. The first exploration used the input difference $0x2800/0010$ of the best 5-round difference path to train the $ND^{cp}_{1\_r}$, in contrast to $0x0040/0000$ employed in [1]. However, the accuracy achieved with $0x2800/0010$, 75.85%, is lower than 92.9% acquired with $0x0040/0000$. The second exploration utilized the differences of the ciphertext pairs to train $ND^{cd}_{1\_r}$. The accuracy of the newly developed 5/6/7-

round neural distinguishers is 90.6%/75.4%/58.3%, which is slightly worse than that of $ND^{cp}_{1\_r}$.

Then, four experiments (Experiments A, B, C and D) were conducted to analyze the ciphertext pairs. Experiment A aimed to assess whether the $ND^{cp}_{1\_5}$ was more likely to recognize the 5-round ciphertext difference with a higher probability as D_r. However, the results showed otherwise, contrary to our initial expectation. After Experiment A, Experiment B investigated the capacity of $ND^{cp}_{1\_5}$ to effectively identify the 5-round ciphertext pairs characterized by a strong probability in their 3/4-round differences. The results of the experiments indicated that $ND^{cp}_{1\_5}$ can adeptly recognize these pairs with an accuracy of approximately 100%.

In addition, they carried out an examination of the bias displayed by the 3/4-round differential bits ($TD_{3/4}$) and noted a strong dependence of $ND^{cp}_{1\_5}$ on $TD_{3/4}$ in Experiment C. To verify the capability of $ND^{cp}_{1\_r}$ in identifying truncated differences, a 2-round neural distinguisher was trained with TD_3 in Experiment D. The accuracy of this 2-round neural distinguisher reached 96.57%, affirming its ability in identifying truncated differences.

From a machine learning perspective, their objective was to investigate the possibility of replacing $ND^{cp}_{1\_r}$ with a hybrid strategy inspired by both differential cryptanalysis and machine learning. To accomplish this objective, they delved into three essential components of the neural network: the convolution layer, the 10-layer residual blocks, and the MLP block. The initial convolution layer transform the input (C_l, C_r, C'_l, C'_r) into $(\Delta L, \Delta V, V_0, V_1)$, alongside a linear amalgamation of these elements, where $(\Delta L, \Delta V, V_0, V_1) = (C_l \oplus C'_l, C_l \oplus C_r \oplus C'_l \oplus C'_r, C_l \oplus C_r, C'_l \oplus C'_r)$. The 10-layer residual blocks in the middle was replaced with the Masked Output Distribution Table (M-ODT). The final MLP block was substituted with a non-neuronal classifier, the Light Gradient Boosting Machine (LGBM) in [18]. As a result, they achieved a non-neuronal model with an accuracy only 0.6% lower than that of the $ND^{cp}_{1_5}$.

Besides investigating the inner workings of $ND^{cp}_{1\_r}$, Benamira et al. also introduced a technique to enhance the accuracy of the neural distinguisher by employing a batch of ciphertext inputs.

3. The Interpretability of the Neural Distinguisher with a ciphertext difference ($ND^{cd}_{1\_r}$)

Although Benamira et al. utilized the ciphertext differences to train $ND^{cd}_{1\_r}$ in [5], they did not delve deeply into its underlying principles. In this section, we aim to complement this investigation. First, we construct the D_r and $ND^{cd}_{1\_r}$ for Speck32/64, employing the difference $0x0040/0000$ as detailed in [1,5]. The accuracy (Acc), true positive rate (TPR), and true negative rate (TNR) for $r \in 5, 6, 7$, are presented in Table 2. Notably, these metrics demonstrate a remarkable similarity, with differences not surpassing 0.01. This observation prompts an intriguing question: *Does $ND^{cd}_{1\_r}$ exploit the same features as D_r to effectively distinguish the ciphertext differences?*

To address this inquiry, we further investigate the potential of $ND^{cd}_{1\_r}$ constructed with the input difference of best r-round differential path following the work in [5]. The results shown in Table 3 indicate that this approach can only yield worse results for $ND^{cd}_{1\_r}$, same to $ND^{cp}_{1\_r}$.

Then Experiment A and Experiment B are conducted to study the correlation between $ND^{cd}_{1\_r}$ and r-round ciphertext difference in Section 3.2. The outcomes of these experiments reveal that $ND^{cd}1\_r$ exhibits a threshold interval denoted as τ (as defined in Definition 1). Specifically, for $ND^{cd}_{1\_r}, r \in 5, 6, 7$, the corresponding τ values are $[2^{-35}, 2^{-30}]$, $[2^{-37}, 2^{-32}]$, and $[2^{-41}, 2^{-36}]$, respectively.

In addition to the r-round ciphertext difference, we also explore the $(r-1)$-round and $(r-2)$-round (truncated) difference in Section 3.3 and 3.4. These studies illustrate that $ND^{cd}_{1\_r}$ can effectively recognize the r-round ciphertext differences that have a high differential probability or bit bias in the last two rounds, similar to the case of $ND^{cp}_{1\_r}$.

Table 2 The Acc, TPR and TNR of the D_r and $ND^{cd}_{1\_r}$ for Speck32/64.

$Type$	$r=5$			$r=6$			$r=7$		
	Acc	TPR	TNR	Acc	TPR	TNR	Acc	TPR	TNR
D_r	0.911	0.877	0.947	0.758	0.680	0.837	0.591	0.543	0.640
$ND^{cd}_{1\_r}$	0.907	0.867	0.946	0.755	0.670	0.836	0.586	0.536	0.635

3.1 Choice of Input Difference

To evaluate whether the $ND^{cd}_{1\_r}$ can achieve improved accuracy with the input difference of the best r-round differential path, we initially employ the Mixed Integer Linear Programming (MILP) technique outlined in [14] to search for all the $5/6/7$-round differential paths with the highest probability for Speck32/64. There are two distinct $5/7$-round differential paths with the best probability 2^{-9} or 2^{-18}, respectively, and one 6-round differential path with the best probability 2^{-13}. Subsequently, we train the $ND^{cd}_{1\_r}$ using the input differences of each differential path. The accuracy of all these neural distinguishers is lower than that trained with the $0x0040/0000$, as shown in Table 3. Specifically, when the number of rounds is 7, there is no usable neural distinguisher. These results collectively indicate that $ND^{cd}_{1\_r}$ can not achieve superior accuracy with the input difference of the best r-round differential path, much like the observed behavior of $ND^{cp}_{1\_r}$.

3.2 The r-round Differences

The pure differential distinguisher D_r distinguishes the various ciphertext differences according to their probability with Formula (2). Consequently, for D_r, a higher probability associated with a ciphertext difference indicates stronger evidence that it originates from the encryption cipher. If $ND^{cd}_{1\_r}$ distinguishes the ciphertext difference using the same features as D_r, its recognition ability for the different ciphertext

Table 3 The best $5/6/7-$round differential paths and accuracy of $ND^{cd}_{1\_r}$ constructed with their input difference for Speck32/64.

r	$log^{DP(\alpha\to\beta)}$	α	β	Acc
5	-9	$0x2800/0010$	$0x850a/9520$	0.750
		$0x0211/0a04$	$0x8000/840a$	0.561
6	-13	$0x0211/0a04$	$0x850a/9520$	0.513
7	-18	$0x0a20/4205$	$0x850a/9520$	Fail
		$0x0a60/4205$	$0x850a/9520$	Fail

differences also aligns with this criterion.

Experiment A. Inspired by this, we analyze the relationship between the score and frequency distribution of the ciphertext differences. In this context, frequency denotes the occurrences of the difference, while the score corresponds to the output of sigmoid activation function, $S(x) = \frac{1}{1+e^{-x}} \in [0, 1]$, at last layer of the neural network.

We use 10^7 plaintext pairs with the plaintext difference $\alpha = 0x0040/0000$ to conduct this experiment for the 5-round Speck32/64, and yield 6,373,162 unique ciphertext differences, including 1,021,347 instances with a frequency greater than 1, and 5,351,815 instances with a frequency of 1. Notably, when the frequency of the ciphertext difference exceeds 1, its score consistently approached 1. This indicates that $ND^{cd}_{1\_5}$ can effective recognizes these differences. Conversely, when the frequency equals 1, the scores of some differences fall below 0.5, accounting for approximately 24.5%.

Experiment B. The probability of ciphertext differences with the frequency 1 approximate 10^{-7} (about $2^{-23.25}$), while the probability for differences with the frequency greater than 1 would surpass it. D_r labels the ciphertext differences with probability greater than 2^{32} as 1 according to Formula 2. Does $ND^{cd}_{1\_r}$ do the same?

For the 1,021,347 ciphertext differences with a frequency greater than 1, it holds true. Consequently, our attention is directed towards the remaining 5,351,815 ciphertext differences with frequency 1. Given the enormity of studying all these ciphertext differences, we randomly selected 10^4 differences for analysis. These differences are categorized into two groups based on whether their scores surpass 0.5, this is also the criterion employed by $ND^{cd}_{1\_5}$ for classifying the ciphertext differences. These differences can be divided into the following three categories:

1. For differences with a probability greater than 2^{-29}, $ND^{cd}_{1\_5}$ can fully identify them.
2. For differences with a probability less than 2^{-35}, $ND^{cd}_{1\_5}$ is unable to identify them.
3. For differences with a the probability in $[2^{-35}, 2^{-29}]$, $ND^{cd}_{1\_5}$ can only identify a part of them.

Figure 1 provides the results for 200 representative differences for the 5-round Speck32. Among these 200 differences, half of them have scores exceeding 0.5, while the remaining half do not surpass this threshold.

Definition 1: Threshold interval τ: For convenience, we defined $[2^{-35}, 2^{-29}]$ as the τ of $ND^{cd}_{1\_5}$. Similarly, the τ of D_r can be expressed as $[2^{-n}, 2^{-n}]$, i.e. 2^{-n}. For example, it is 2^{-32} for Speck32/64.

In order to ascertain the generality and universality of this phenomenon, we extend our experimentation to $ND^{cd}_{1\_6}$ and $ND^{cd}_{1\_7}$. A similar pattern is observed for $ND^{cd}_{1\_6}$ and $ND^{cd}_{1\_7}$, their threshold intervals are $[2^{-38}, 2^{-32}]$ and $[2^{-41}, 2^{-36}]$, respectively.

These experiments indicate that the τ for $ND^{cd}_{1\_r}$ varies for different rounds r, and it diminishes as r increases, which is consistent with the fact that an increase in the number of rounds results in a decrease in differential probability. However, for D_r, its bound is consistently fixed to 2^{-n}. In conclusion, although D_r and $ND^{cd}_{1\_r}$ exhibit similar values for Acc, TPR, and TNR, their inner principles are different.

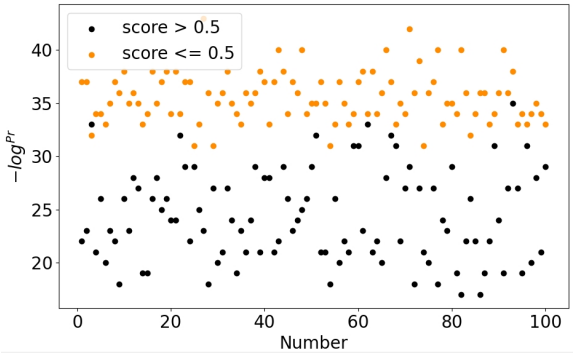


Fig. 1 The $differential\ probability$ and $Score$ of 200 representative differences for $ND^{cd}_{1\_5}$.

3.3 The $(r-1)$-round or $(r-2)$-round Differences

Experiment C. Following Experiment B in [5], this section employs $ND^{cd}_{1\_5}$ as a representative instance to probe the capacity of $ND^{cd}_{1\_r}$ to identify the differences in rounds $(r-1)$ and $(r-2)$. The Experiment B in previous section indicates that $ND^{cd}_{1\_5}$ can effectively identify all the differences with a probability exceeding 2^{-29}. Therefore, if the 5-round difference has a high probability in the round 3/4, there is a significant chance that $ND^{cd}_{1\_5}$ can identify this 5-round difference, unless the differential probability in the last two rounds or one round is very low, causing the 5-round probability fall below 2^{-29}.

To verify this conjecture, we employ 10^7 plaintext pairs with the difference $0x0040/0000$ to calculate the frequency of all the 3/4-round ciphertext differences, denoted as $Diff_{3/4}$. The number of differences we obtain in $Diff_{3/4}$ is 41,515 and 1,705,584, respectively. These differences are organized in a descending order based on their frequency. Subsequently, 10^7 regenerated plaintext pairs are utilized to extract the 5-round differences whose 3/4-round differences belong to $Diff_{3/4}$. Finally, $ND^{cd}_{1\_5}$ is applied to evaluate the remaining 5-round differences, yielding an accuracy of 95.33%/86.94%. The result of $r = 3$ is consistent with the TPR of $ND^{cd}_{1\_5}$ in Table 2 (0.867). This observation is consistent with the findings for $ND^{cp}_{1\_5}$ shown in [5].

3.4 The $(r-1)$-round or $(r-2)$-round Truncated Differences

Experiment D. In addition to focusing solely on the ciphertext difference, we also examine the bias of the differential bits following the Experiment C in [5]. To assess the ability of $ND^{cd}_{1\_r}$ to identify the specific bits in the $(r-1)/(r-2)$-round differences, we initially derive the differential set B (with a score of less than 0.1) and G (with a score exceeding 0.9) from the 10^7 5-round ciphertext pairs. Subsequently, we count the bit bias of different bits. The 3/4-round truncated difference $TD_{3/4}$ extracted from B and G is
TD_3 : 01 * * * *000 * * * * * 01 * * * * * *000 * * * * * 01
TD_4 : 00 * * * * * * * * * * * * * *01 * * * * * * * * * * * * * * ,
where $*$ denotes an uncertain binary bit.

Then we regenerate the 10^6 plaintext pairs to derive the 5-round differences whose 3/4-round differences satisfying $TD_{3/4}$. These differences are then evaluated with $ND^{cd}_{1\_5}$, resulting in accuracy of 98.4% and 97.1%, respectively. This demonstrates that $ND^{cd}_{1\_5}$ possesses the capability to identify specific bits of the $(r-1)/(r-2)$-round differences , in a manner analogous to that of $ND^{cp}_{1\_5}$.

Experiment E. In order to assess the ability of neural distinguishers to detect truncated differences, a 2/3-round truncated differential neural distinguisher is trained utilizing TD_3. Firstly, 10^7 plaintext pairs are generated, with half of the pairs satisfying TD_3 and the other half not. Subsequently, the ciphertext differences obtained by encrypting these plaintext pairs are employed to train the 2/3-round truncated differential neural distinguishers with the same neural network architecture and parameters in Section 2.4. The accuracy of the 2/3-round truncated differential neural distinguisher is 95.88%/65.67%. This indicates that the neural distinguishers using ciphertext differences can also effectively identify the truncated differences, similar to the neural distinguishers employing ciphertext pairs as shown in [5].

4. Comparing the Neural Distinguisher with a Ciphertext Difference or a Ciphertext Pair

It is evident that $ND^{cp}_{1\_r}$ and $ND^{cd}_{1\_r}$ exhibit strikingly similar behavior from the experiments in the previous section. For instance, the effects of $(r-1)$-round or $(r-2)$-round (truncated) differences are identical for both of them. Based on the fact that $ND^{cd}_{1\_r}$ possesses a τ, we first conduct an investigation to determine whether $ND^{cp}_{1\_r}$ has a τ like $ND^{cd}_{1\_r}$ in Experiment A. The results confirm this hypothesis, revealing that the τ of $ND^{cp}_{1\_r}$ is very similar or identical to that of $ND^{cd}_{1\_r}$.

The only difference between $ND^{cp}_{1\_r}$ and $ND^{cd}_{1\_r}$ is their sample format: the former is trained with ciphertext pairs, whereas the latter is trained with their differences. To enhance our understanding of their connections and differences from a sample-based perspective, we conduct Experiment B and Experiment C. Experiment B is designed to confuse the feature relationship of ciphertext pairs by imposing random values, while Experiment C aims to discern the importance of different bits for both $ND^{cp}_{1\_r}$ and $ND^{cd}_{1\_r}$. The results of these two experiments reveal that $ND^{cp}_{1\_r}$ not only captures the differences between ciphertext pairs, but also identifies a specific correlation among certain ciphertext bits between the left and right halves of the ciphertext, especially the significant bits.

4.1 A Cryptanalysis Perspective

Experiment A. To determine whether $ND^{cp}_{1\_r}$ has a threshold interval τ similar to $ND^{cd}_{1\_r}$, we employ $ND^{cp}_{1\_r}$ to distinguish the ciphertext pairs with various differential probability, following the Experiment B in Section 3.2. Here, we consider all differences, regardless of their frequency, since the neural distinguisher can effectively identify differences with frequency exceeding 1.

The τ of $ND^{cp}_{1\_5}$ is $[2^{-36}, 2^{-30}]$ (according to the Definition 1), as shown in Figure 2. Notably, this range closely mirrors that of $ND^{cd}_{1\_5}$. A similar experiment is also conducted on $ND^{cp}_{1\_6}$ and $ND^{cp}_{1\_7}$. Their threshold intervals τ are $[2^{-37}, 2^{-32}]$ and $[2^{-41}, 2^{-36}]$, respectively. It aligns with that of $ND^{cd}_{1\_6}$ and $ND^{cd}_{1\_7}$. Based on these findings, we can confidently conclude that $ND^{cp}_{1\_r}$ also possesses a threshold interval τ that is either identical or very similar to that of $ND^{cd}_{1\_r}$.

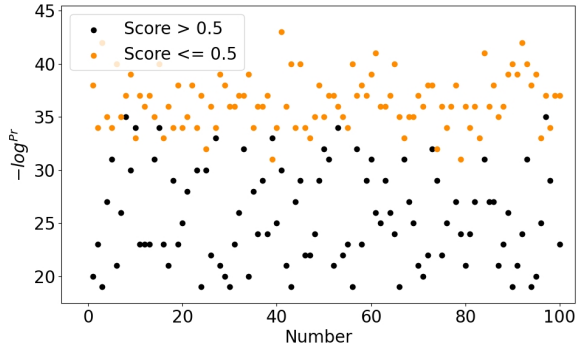


Fig. 2 The $differential\ probability$ and $Score$ of 200 representative ciphertext pairs for $ND^{cp}_{1\_5}$.

4.2 A Feature Confusion Perspective

Experiment B. To clarify the factors contributing to the superior accuracy of $ND^{cp}_{1\_r}$ over $ND^{cd}_{1\_r}$, we carry out the following investigations. Firstly, we train them for rounds 5, 6 and 7. Then we retrain them with the confused ciphertext pairs in the experiments $Exp.B-1$, $Exp.B-2$ and $Exp.B-3$. $Exp.B-1$ confuses the ciphertext pairs with the same random values to break the relation between the ciphertext bits but ensure the ciphertext difference remains constant. For a ciphertext pair $((C_{0,l}, C_{0,r}), (C_{1,l}, C_{1,r}))$, $(C'_{i,l}, C'_{i,r})$ is $(C_{i,l} \oplus R_0, C_{i,r} \oplus R_1)$, $i \in [0,1]$, where R_0 and R_1 are the different random values. In contrast, $Exp.B-2$ and $Exp.B-3$ only confuse the left or right half of the ciphertext, i.e. $C'_{i,l} = C_{i,l} \oplus R_0$ or $C'_{i,r} = C_{i,r} \oplus R_1$, $i \in [0,1]$. The results of these three experiments are summarized in Table

4. Notably, the accuracy of the neural distinguishers in $Exp.\ B-1$, $Exp.\ B-2$ and $Exp.\ B-3$ is equal to that of $ND^{cd}_{1\_r}$, indicating that these distinguishers only learn the ciphertext difference from the confused ciphertext pairs.

Comparing the results of $Exp.\ B-1$ and $Exp.\ B-2$ ($Exp.\ B-3$), we observe that $ND^{cp}_{1\_r}$ fails to extract any valuable features from $(C_{0,r}, C_{1,r})$ or $(C_{0,l}, C_{1,l})$ apart from their differences. This suggests that the additional features learned by $ND^{cp}_{1\_r}$ may be the relationship between the left and right halves of the ciphertext, possibly due to the Feistel structure of Speck32/64. To delve deeper into this, we conduct $Exp.\ B-4$, which confuses the left and right halves of the ciphertext with the same random values, i.e. $(C'_{i,l}, C'_{i,r}) = (C_{i,l} \oplus R_0, C_{i,r} \oplus R_0), i \in [0,1]$. The accuracy of the neural distinguisher in this experiment is equal to that of $ND^{cp}_{1\_r}$, suggesting a certain correlation between the left and right halves of the ciphertext. In the next section, we will explore this correlation further by examining the feature importance of the different ciphertext bits.

Table 4 The accuracy of the 5/6/7-round neural distinguisher with different data format for Speck32/64. $(C_{0,l}, C_{0,r})$ and $(C_{1,l}, C_{1,r})$ denote the first and second ciphertext, where $C_{0,l}$ and $C_{1,l}$ are the left half of the ciphertext, while $C_{0,r}$ and $C_{1,r}$ are the right half. For the X, X' is equal to X XOR a random value R, such as $C'_{0,l} = C_{0,l} \oplus R$.

$Exp.$	Samples format	$Acc.$		
		$r=5$	$r=6$	$r=7$
$ND^{cp}_{1\_r}$	$(C_{0,l}, C_{0,r}, C_{1,l}, C_{1,r})$	0.926	0.785	0.611
$ND^{cd}_{1\_r}$	$(C_{0,l} \oplus C_{1,l}, C_{0,r} \oplus C_{1,r})$	0.907	0.755	0.586
$B-1$	$(C'_{0,l}, C'_{0,r}, C'_{1,l}, C'_{1,r})$	0.907	0.755	0.586
$B-2$	$(C'_{0,l}, C_{0,r}, C'_{1,l}, C_{1,r})$	0.907	0.755	0.586
$B-3$	$(C_{0,l}, C'_{0,r}, C_{1,l}, C'_{1,r})$	0.907	0.755	0.586
$B-4$	$(C'_{0,l}, C'_{0,r}, C'_{1,l}, C'_{1,r})$	0.926	0.785	0.611

The $Exp.\ B-1$ confused the $C_{i,l}$ and $C_{i,r}$, $i \in [0,1]$, with the different random values.
The $Exp.\ B-4$ confused the $C_{i,l}$ and $C_{i,r}$, $i \in [0,1]$, with the same random value.

4.3 A Feature Importance Perspective

Experiment C. Permutation Feature Importance [19] (PFI) is a model-independent method to measure the feature importance based on the extent of model score reduction upon feature permutation. For $ND^{cd}_{1\_r}$, we directly permute its ciphertext bits, denoted as per_cd. For $ND^{cp}_{1\_r}$, we conduct three different experiments to observe the patterns it learns: the initial two experiments only permute the $i-$th bit of the left or right (first or second) ciphertext, denoted as per_cp_l or per_cp_r, while the third experiment permute the $i-$th bit of both ciphertexts, denoted as per_cp_lr. To ensure the reliability of the experiment and fairly compare the different neural distinguishers, we repeat the evaluation 20 times and consider the weight of the average reduced accuracy over original accuracy as the final feature importance. The comprehensive procedure is given in Algorithm 3.

The criteria used to assess the feature importance of various bits for $ND^{cd}_{1\_r}$ and $ND^{cp}_{1\_r}$ are as follows:

1. If the feature importance of the $i-$th bit ($Fi[i]$) is greater than 0 for the experiment per_cd, $ND^{cd}_{1\_r}$ learns the features from the $i-$th bit ciphertext difference.
2. For the experiment per_cp_l or per_cp_r, if $Fi[i] > 0$, $ND^{cp}_{1\_r}$ learns the features from the $i-$th bit ciphertext difference.
3. If $Fi[i]$ in experiment per_cp_lr surpasses that in per_cp_l or per_cp_r, the $i-$th bit ciphertext contributes extra information to $ND^{cp}_{1\_r}$ beyond just its difference. We call these bits as *significant bits*.
4. In the all experiments, the larger the value of $Fi[i]$, the higher the importance of the $i-$th bit.

Algorithm 3 Calculating feature importance of each ciphertext bit.

Input: Neural network Net
Output: Feature importance $Fi[n]$, where n is the length of the ciphertext.
for $i \leftarrow 0, n$ **do**
  $Acc_{sum} = 0$
  for $j \leftarrow 0, 20$ **do**
    Generate the original samples X and labels Y.
    $X_{perm} \leftarrow$ Randomly permuting the $i-$th bit for X.
    $Acc_{sum} \leftarrow Acc_{sum} + Net.evaluate(X, Y) - Net.evaluate(X_{perm}, Y)$
  end for
  $Fi[i] = \frac{Acc_{sum}}{20}$
end for
return Fi

We apply this algorithm to the 5/6/7-round Speck32/64 with 10^6 samples. As expected, the bit importance of per_cp_l or per_cp_r are consistent, since the first and second ciphertexts are unordered and exchangeable. Moreover, the bits learned by $ND^{cd}_{1\_r}$ and $ND^{cp}_{1\_r}$ are identical. In most cases, $ND^{cp}_{1\_r}$ can acquire more knowledge from the different ciphertext bits. However, sometimes $ND^{cd}_{1\_r}$ relies more heavily on certain bits than $ND^{cp}_{1\_r}$, such as the 2nd and 16th bit for rounds 5, as shown in Figure 3. This could be attributed to the simpler data format, which enables the neural network to effectively gain more information from the ciphertext bits.

For $ND^{cp}_{1\_r}$, it not only captures the difference of ciphertext pairs, but also discerns the relationships among the ciphertext bits, particularly the significant bits such as the 2nd, 3rd, 18th and 19th bit for 5 rounds. These significant bits exist in the corresponding left and right parts of the ciphertext, such as the 2nd bit and its corresponding 18th bit. This further validates our findings in the Experiment B of Section 4.2.

5. Exploring and Enhancing the Neural Distinguisher with Multiple Ciphertext Pairs

The experiments in Section 4 reveal that $ND^{cp}_{1\_r}$ predominantly learns the ciphertext differences and some subtle

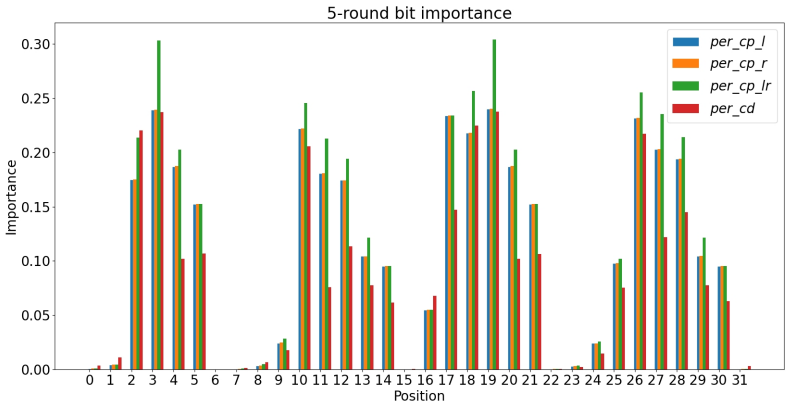


Fig. 3 The bit importance of 5-round neural distinguisher.

relationships among certain ciphertext bits, especially the significant bits. For $ND^{cp}_{k\_r}$, its sample consists of k ciphertext pairs. Consequently, the information that $ND^{cp}_{k\_r}$ can extract from its sample may be the difference of k ciphertext pairs and relationships between the ciphertexts themselves or ciphertext pairs.

In order to figure out the features that $ND^{cp}_{k\_r}$ learned, we conduct three different experiments (Experiment A, B and C) to explore its intrinsic principles. Through studying the differential distributions of multiple ciphertext pairs in Experiment A, we discover that the average differential probability of the multiple ciphertext pairs is the paramount criterion for $ND^{cp}_{k\_r}$ to distinguish the samples. This is the reason why $ND^{cp}_{k\_r}$ can achieve better accuracy as k increases. As the number of ciphertext pairs in the sample increases, those anomalous differential probabilities are neutralized, resulting in the average differential probabilities of positive and negative samples stabilizing within the distinct interval.

In addition, there are other features that allow $ND^{cp}_{k\_r}$ to recognize some samples with abnormal average differential probabilities, and the most intuitive one is the bit relationship between the ciphertexts or ciphertext pairs. Especially when the number of rounds is small, this feature is more obvious. For example, when the number of rounds is 5, the neural network can learn useful features from the sample (C'_0, C_1, C'_2, C_3) in $Exp.\ B-6$. Furthermore, we find that the features between ciphertext pairs are fragile, this is also the reason why Gohr et al. can use the combined scores of multiple ciphertext pairs to achieve competitive results with $ND^{cp}_{k\_r}$ in [20].

Based on the fact that $ND^{cp}_{k\_r}$ mainly relies on the average differential probability of the k ciphertext pairs for sample recognition, we propose a approach to construct the neural distinguisher with higher accuracy by generating negative samples using a fixed difference instead of a random one. With this method, we obtain the neural distinguishers with higher accuracy for both 7-round Speck32/64 and 9-round Simon32/64.

5.1 A Cryptanalysis Perspective

Experiment A. Following Experiment A in Section 4.1, we first analyze the differential probability distribution of the samples for $ND^{cp}_{2\_5}$. In contrast to the previous experiment, the sample of $ND^{cp}_{2\_5}$ contains 2 ciphertext pairs instead of one, prompting us to conduct two sub-experiments to observe their differential probability. The first sub-experiment focus on the individual differential probability of the two ciphertext pairs, while the second sub-experiment consider their average differential probability.

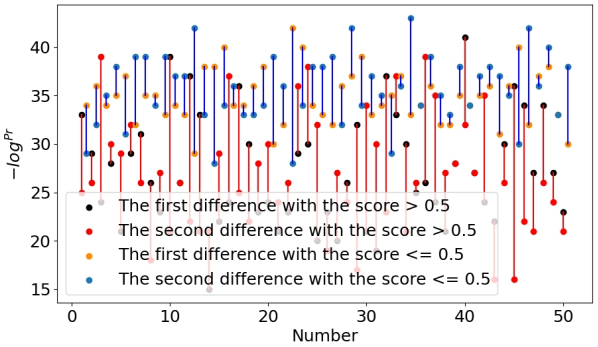


Fig. 4 The *differential probability* of the 100 samples for the $ND^{cp}_{2\_5}$.

The differential probability of 100 randomly selected samples is given in Figure 4, with half of the samples having a $score > 0.5$ and the remaining half having $score \leq 0.5$. It is obvious that the majority of samples with a score greater than 0.5 exhibit a higher differential probability compared to those with a score ≤ 0.5. However, certain samples deviate from this distribution. For instance, the differential probabilities of the two ciphertext pairs in the 33rd sample with the score > 0.5 are 2^{-33} and 2^{-37}, whereas for the 38th sample with a score ≤ 0.5, they are 2^{-32} and 2^{-33}. Furthermore, the τ of $ND^{cp}_{1\_5}$ is $[2^{-35}, 2^{-29}]$, indicating that $ND^{cp}_{1\_5}$ cannot identify the ciphertext pairs with a differential probability less than 2^{-35}. However, $ND^{cp}_{2\_5}$ can recognize some samples where one of the two ciphertext pairs has a differential probability below 2^{-35}. This implies an interaction between the ciphertext pairs. Therefore, we further investigate their average differential probability distributions.

The average differential probability of 200 random samples, divided into two groups based on their scores ($score > 0.5$ or $score \leq 0.5$), for $ND^{cp}_{2\_5}$ is presented in Figure 5. It is evident that 2^{-32} serves as an effective threshold for determining whether the sample score exceeds 0.5 for

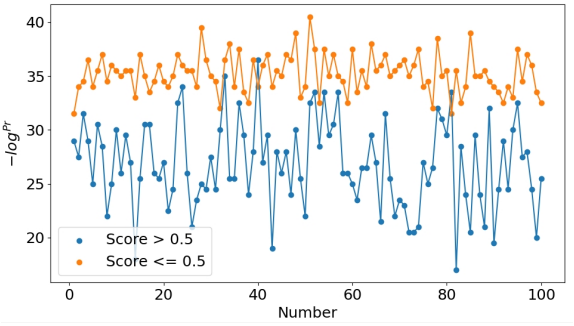


Fig. 5 The average differential probability of the 200 samples for $ND^{cp}_{2\_5}$.

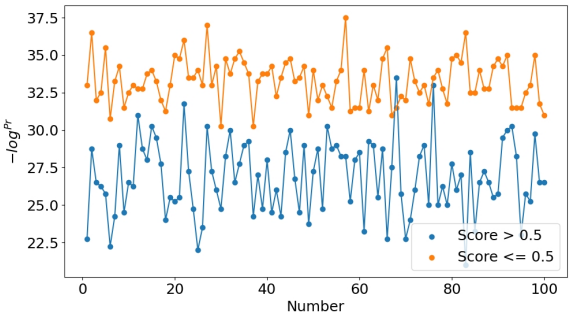


Fig. 6 The average differential probability of the 200 samples for $ND^{cp}_{4\_5}$.

the majority of samples. To further assess the impact of average differential probability on $ND^{cp}_{k\_r}$, we also performed this experiment for $ND^{cp}_{4\_5}$. Its threshold is approximately 2^{-30}, as shown in Figure 6. Compared to $ND^{cp}_{2\_5}$, $ND^{cp}_{4\_5}$ can achieve better sample scoring and exhibit a more distinct boundary. As the number of ciphertext pairs in a sample increases, the average differential probability of these ciphertext pairs steadily converges towards values within the range $[2^{-31}, 2^{-22}]$. This convergence is one of the reasons why $ND^{cp}_{k\_r}$ can achieve higher accuracy as k increases. Moreover, the features learned by $ND^{cp}_{k\_r}$ encompass not only the average differential probability of the k ciphertext pairs but also other relations within the ciphertexts. Consequently, some samples with the abnormal average differential probability can also be identified. For example, the 40th sample in Figure 5 and the 69th sample in Figure 6. We further examine what these relationships are in the following two sections.

5.2 A Feature Confusion Perspective

Experiment B. To investigate the features learned by $ND^{cp}_{2\_r}$, we conduct a series of experiments outlined in Table 5 to observe the impact of different features on $ND^{cp}_{2\_r}$. In $Exp.\ B-1$, $Exp.\ B-2$, and $Exp.\ B-3$, we confuse the bit relationships in the ciphertext pairs while preserving their differences intact. Specifically, $Exp.\ B-1$ only adds confusion to the first ciphertext pair, whereas $Exp.\ B-2$ and $Exp.\ B-3$ introduce confusion to both ciphertext pairs. The main difference between $Exp.\ B-2$ and $Exp.\ B-3$ is the random value used for confusion. In $Exp.\ B-2$, the two ciphertext pairs are confused with the same random values, whereas in $Exp.\ B-3$, different random values are used. The purpose of $Exp.\ B-2$ and $Exp.\ B-3$ is to ascertain whether $ND^{cp}_{2\_r}$ can discern the relationship among the two ciphertext pairs.

Table 5 The accuracy of the 5/6/7-round $ND^{cp}_{2\_r}$ with different data format for Speck32/64. (C_0, C_1) and (C_2, C_3) denote the two ciphertext pairs and the C'_i is equal to C_i XOR a random value R, i.e. $C'_0 = C_0 \oplus R$.

$Exp.$	Samples format	$Acc.$		
		$r=5$	$r=6$	$r=7$
$ND^{cp}_{1\_r}$	(C_0, C_1)	0.926	0.785	0.611
$ND^{cd}_{1\_r}$	$(C_0 \oplus C_1)$	0.907	0.755	0.586
$ND^{cp}_{2\_r}$	(C_0, C_1, C_2, C_3)	0.978	0.871	0.647
$ND^{cd}_{2\_r}$	$(C_0 \oplus C_1, C_2 \oplus C_3)$	0.968	0.842	0.618
$B-1$	(C'_0, C'_1, C_2, C_3)	0.972	0.858	0.635
$B-2$	(C'_0, C'_1, C'_2, C'_3)	0.968	0.842	0.618
$B-3$	(C'_0, C'_1, C'_2, C'_3)	0.968	0.842	0.618
$B-4$	(C'_0, C_1, C_2, C_3)	0.926	0.785	0.611
$B-5$	(C'_0, C'_1, C_2, C_3)	0.926	0.785	0.611
$B-6$	(C'_0, C_1, C'_2, C'_3)	0.912	0.755	0.586
$B-7$	(C'_0, C_1, C'_2, C_3)	0.820	0.501	0.501

The C'_0 and C'_1 are confused with the same random value to keep their differences invariant.
The two ciphertext pairs are confused with the same random value to leaves the relations between ciphertext differences unchanged.
The two ciphertext pairs are confused with the different random values to confuse the relations between ciphertext differences.
The C'_0 and C'_1 are confused with the different random values to completely destroys the features in the first ciphertext pair.

As expected, the accuracy of $Exp.\ B-1$ surpasses that of $ND^{cd}_{2\_r}$ and a little less than that of $ND^{cp}_{2\_r}$, since the first confused ciphertext pair (C'_0, C'_1) can only contribute its difference. The accuracy of the neural distinguishers in $Exp.\ B-2$ and $Exp.\ B-3$ is consistent with that of $ND^{cd}_{2\_r}$, suggesting that these neural distinguisher only learned the difference of the ciphertext pairs. When we only confuse the first ciphertext in $Exp.\ B-4$ or the first ciphertext pair in $Exp.\ B-5$ with the different random values, the accuracy of their neural distinguishers matches to that of $ND^{cp}_{1\_r}$, indicating the neural distinguisher can not extract valuable information from the first ciphertext pair. By comparing $Exp.\ B-4$ and $Exp.\ B-5$, it is evident that the neural distinguisher does not learn useful features from the C_1 in $Exp.\ B-4$.

Furthermore, building upon the findings of $Exp.\ B-4$, we proceed to confuse the second ciphertext pair while preserving their difference in $Exp.\ B-6$. As a result, the accuracy of the neural distinguishers for the rounds 6 and 7 is equal to that of $ND^{cd}_{1\_6}$ and $ND^{cd}_{1\_7}$. However, for round 5, its accuracy surpasses that of $ND^{cd}_{1\_5}$, which indicates that the neural network acquires some knowledge from the C_1. This can be attributed to the fact that when the number of

rounds is 5, the ciphertext contains valuable information that can be easily learned by the neural network. This is corroborated by $Exp.\ B-7$, when we confuse the first ciphertext in each ciphertext pair, the neural network can only extract useful information for $r = 5$. This information may be the distribution information of the ciphertext itself.

For the $Exp.\ B-6$, the neural network learn the features from the ciphertext C_1 for $r = 5$. However, in $Exp.\ B-4$, C_1 does not enhance the accuracy of the neural distinguisher. This discrepancy arises due to the fact that the second ciphertext pair in $Exp.\ B-6$ is confused, resulting in the destruction of its ciphertext features. Conversely, in $Exp.\ B-4$, the second ciphertext pair remains intact, preserving these features. Consequently, in $Exp.\ B-6$, the ciphertext features provided by C_1 are duplicated by C_2 and C_3. In contrast, in $Exp.\ B-6$, the second ciphertext pair only provides their differences, making the ciphertext features provided by C_1 valuable. However, it is important to note that these features are inherently fragile and only effective for round 5.

5.3 A Feature Importance Perspective

Experiment C. Following Experiment C in Section 4.3, we conduct 5 different experiments ($Exp.\ fi\_0$ to $Exp.\ fi\_4$) to analyze the feature importance of different ciphertext bits for $ND^{cp}_{2\_r}$ using Algorithm 3. For a given sample (C_0, C_1, C_2, C_3), the $Exp.\ fi\_0$ and $Exp.\ fi\_1$ permute the $i-$th bit of the C_0 or (C_0, C_1) in the first ciphertext pair, respectively, whereas the $Exp.\ fi\_2$, $Exp.\ fi\_3$ and $Exp.\ fi\_4$ separately permute the $i-$th bit of the (C_0, C_2), (C_0, C_1, C_2) and (C_0, C_1, C_2, C_3).

When we only permute the $i-$th bit of the first ciphertext pair in $Exp.\ fi\_0$ and $Exp.\ fi\_1$, the feature importance of all the ciphertext bits is less than 0.08. However, when we permute the $i-$th bit of both ciphertext pairs in $Exp.\ fi\_2$, $Exp.\ fi\_3$ and $Exp.\ fi\_4$, the feature importance of certain bits, such as the 2nd and 19th bits, increase significantly. This further supports the findings from $Exp.\ B-4$ and $Exp.\ B-5$ in Table 5 that a correct ciphertext pair in the sample can enables the neural distinguisher to effective recognition this sample. For the $Exp.\ fi\_2$, $Exp.\ fi\_3$ and $Exp.\ fi\_4$, their bit importance can be categorized into three cases:

1. If their feature importance increases sequentially, there are features between the $i-$th bit ciphertext and the other ciphertext bits, such as the 2nd, 3rd, 4th, 18th, 19th, and 20th bits.
2. If their feature importance remains the same, the $i-$th bit ciphertext only provides their differences. For example, the 5th, 10th, 21st, 25th, 26th bits.
3. If their feature importance decreases sequentially, there are relationships among all the $i-$th bits ciphertexts in the sample, such as the 16th, and 17th bit.

In addition, by comparing the results in Figure 3 and Figure 7, we observe that the bits with high importance for $ND^{cp}_{1\_5}$ and $ND^{cp}_{2\_5}$ are identical, and their relative importance remains consistent.

Based on these observations, it can be inferred that, even though $ND^{cp}_{k\_r}$ employs multiple ciphertext pairs for sample recognition, its primary reliance lies in the ciphertext pairs themselves (i.e., ciphertext difference and the relations of certain ciphertext bits), the relationships between the ciphertext pairs are extremely fragile. This insight helps clarify why Gohr et al. were able to achieve comparable or superior results for $ND^{cp}_{k\_r}$ in [20] by combining the accuracy of $ND^{cp}_{1\_r}$ on multiple ciphertext pairs.

Table 6 The neural distinguishers for 7-round Speck32/64 and 9-round Simon32/64.

Cipher	Rounds	$k=1$	$k=2$	$k=4$	source
Speck32/64	7	0.611	0.645	0.687	[8][1]
		0.626	0.667	0.713	$Ours$[1]
		–	0.665	0.728	[10][2]
		0.634	0.691	0.764	$Ours$[2]
Simon32/64	9	0.603	0.645	0.666	[8][1]
		0.672	0.704	0.741	$Ours$[1]
		–	0.724	0.810	[10][2]
		0.731	0.805	0.886	$Ours$[2]

Train with $10^7/k$ samples, each containing k $(C_{0,l}, C_{0,r}, C_{1,l}, C_{1,r})$.
Train with 10^7 samples, each containing k $(C_{0,l}, C_{0,r}, C_{1,l}, C_{1,r}, R_0 \oplus R_1)$.

5.4 Enhancing the Neural Distinguisher

Based on the previous research, it is evident that the $ND^{cp}_{k\_r}$ distinguishes the positive samples (the ciphertext pairs obtained by encrypting the plaintext pairs with a specific difference α) and negative samples (the ciphertext pairs derived from encrypting the plaintext pairs with the random difference) primarily based on their differential distributions. For the positive samples, their output difference must align with the differential distribution table of α. However, for the negative samples, their output difference is unrestricted due to the absence of limitations on their plaintext difference. Consequently, there exists a conflict in the differential probability between certain negative and positive samples. This conflict is the underlying reason why $ND^{cp}_{1\_r}$ fails to recognize some ciphertext pairs with differential probabilities in τ. $ND^{cp}_{k\_r}$ improves its accuracy by using the average differential probability of k ciphertext pairs, thereby minimizing the overlap of differential probabilities between positive and negative samples. Building on this, and considering the unbounded nature of negative samples, we could potentially develop a superior neural distinguisher. This improvement would be due to the reduction in conflicts and the increased regularity of the negative ciphertexts.

Inspired by the idea in [21], which utilizes the output difference of the t (≥ 2) input differences for training the neural distinguisher, we present a scheme to enhance the $ND^{cp}_{k\_r}$ by using a fixed difference to generate negative samples instead of a random one, and this new neural distinguisher is named $ND^{cp'}_{k\_r}$. Its construction process is very similarly to that of $ND^{cp}_{k\_r}$, as described in Section 2.4. The necessary

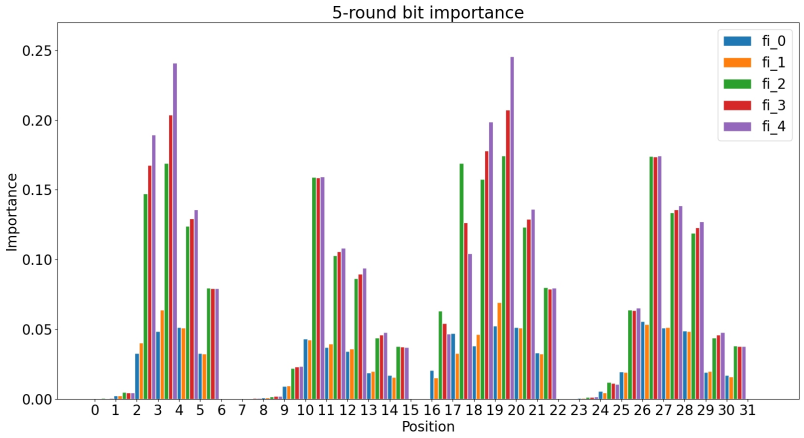


Fig. 7 The bit importance of 5-round neural distinguisher with 2 ciphertext pairs.

modifications are limited to the data generation process and the definition of their corresponding labels. $ND^{cp'}_{k\_r}$ requires two plaintext differences α_0 and α_1 ($\alpha_0 \neq \alpha_1$) to generate positive and negative samples respectively, and the label Y_j of the samples transitions from Equation 2 to Equation 3.

$$Y_j = \begin{cases} 1, \ if \ P_{j,0} \oplus P_{j,1} = \alpha_0, \ j \in [0, k-1] \\ 0, \ if \ P_{j,0} \oplus P_{j,1} = \alpha_1, \ j \in [0, k-1] \end{cases} \tag{3}$$

To validate the effectiveness of our approach, we apply it to the 7-round Speck32/64 and 9-round Simon32/64. It's worth noting that here we're only aiming to demonstrate the superiority of $ND^{cp'}_{k\_r}$ over $ND^{cp}_{k\_r}$, and we're not committed to training the neural distinguishers for more rounds. Since it require a staged approach, such as an 8-round neural distinguisher for Speck32/64 in [1], which would require enormous resources. The α_0 remains as $0x0040/0000$ and $0x0000/0040$ that used in [20], and the α_1 is chosen from the remaining 31 differences with hamming weight 1 for both of them. We only consider the differences with hamming weight 1 here, since they may diffuse more slowly.

We first train $ND^{cp}_{k\_r}$ and $ND^{cp'}_{k\_r}$, $k \in [1, 2, 4]$, with the 10^7 ciphertext pairs for all the 31 α_1 with hamming weight 1 ($\alpha_0 \neq \alpha_1$). † Through these experiments, we find that the accuracy of the majority of $ND^{cp'}_{k\_r}$ is greater than that of $ND^{cp}_{k\_r}$. For the 7-round Speck32/64, the number of the $ND^{cp'}_{k\_7}$, $k \in [1, 2, 4]$, with the accuracy better than or equal to that of $ND^{cp}_{k\_7}$ is 17, 31 and 27. For the 9-round Simon32/64, it is 28, 23 and 26. The best $ND^{cp'}_{k\_r}$, $k \in [1, 2, 4]$, is trained with the $\alpha_1 = 0x0000/8000$ for Speck32/64. For Simon32/64, it is $0x0000/2000$, $0x0000/0008$ and $0x0000/0200$ for $k \in [1, 2, 4]$. The best accuracy is given in Table 6, and the models are provided in our github repository.

Furthermore, we also train the $ND^{cp'}_{k\_r}$ with the new data format used in [10] to illustrate the applicability of our scheme to different data formats. For the Simon32/64, it is $(C_{0,l}, C_{0,r}, C_{1,l}, C_{1,r}, R_0 \oplus R_1)$, where R_i is the right branches of a state after the encryption of $(r-1)$ rounds, i.e., $R_i = (C_{i,r} \lll 8) \wedge (C_{i,r} \lll 1) \oplus (C_{i,r} \lll 2) \oplus C_{i,r}$. For the Speck32/64, it is $(C_{0,l}, C_{0,r}, C_{1,l}, C_{1,r}, R_0, R_1)$, where $R_i = (C_{i,l} \oplus C_{i,r}) \ggg 2$. As in the original paper, we use 10^7 samples and let each sample contain k instances to train the neural distinguisher. This more complex samples make our neural network architecture in Section 2.4 inapplicable, so we adopted the neural network architecture and parameters in [10] here. As expected, this more complex data format yields better accuracy, as shown in Table 6, the corresponding models and evaluation code are available in the github repository.

In addition, we also compare our results with existing results in Table 6. To the best of our knowledge, our accuracy is the best one with the same amount of training data. These experimental results indicate that using a specific difference to generate the negative samples, as opposed to random differences, is an effective approach for enhancing the accuracy of the neural distinguisher.

6. Conclusions

In this paper, we first explore the internal mechanisms of the neural distinguisher constructed using a single ciphertext difference from four aspects: choice of input difference, r-round differences, $(r-1)$-round or $(r-2)$-round difference and truncated difference. Through our analysis, we identify significant similarities between this neural distinguisher and another one trained with a ciphertext pair.

Following this, we conduct a comprehensive comparison of their similarities and differences in the context of differential cryptanalysis, feature confusion, and feature importance. We highlight that the neural network captures the distinct correlation between specific bits of the left and right halves from the ciphertext, especially the significant bits.

After that, our analysis extends to the neural distinguisher with multiple ciphertext pairs. Our investigation indicates that this neural distinguisher recognizes its samples relying heavily on the average differential probability of the ciphertext pairs in the samples. In order to mitigate the conflicts of differential probability between positive and negative samples, we adopt a predetermined difference instead of a random one to generate the negative samples, thereby achieving higher accuracy for the 7-round Speck32 and 9-round Simon32/64.

†The accuracy of all neural distinguishers is listed in *https://github.com/differentialdistinguisher/ND_interpretability*.

Acknowledgments

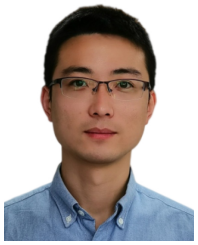
This work is supported by the National Key Research and Development Program of China (2022YFB2701900), the National Natural Science Foundation of China (No. 62072181) and the Shanghai Trusted Industry Internet Software Collaborative Innovation Center.

References

[1] A. Gohr, "Improving attacks on round-reduced speck32/64 using deep learning," Annual International Cryptology Conference, pp.150–179, Springer, 2019.
[2] B. Hou, Y. Li, H. Zhao, and B. Wu, "Linear attack on round-reduced des using deep learning," Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part II 25, pp.131–145, Springer, 2020.
[3] B. Zahednejad and L. Lyu, "An improved integral distinguisher scheme based on neural networks," International Journal of Intelligent Systems, vol.37, no.10, pp.7584–7613, 2022.
[4] L. Lerman, G. Bontempi, and O. Markowitch, "Side channel attack: an approach based on machine learning," Center for Advanced Security Research Darmstadt, vol.29, 2011.
[5] A. Benamira, D. Gerault, T. Peyrin, and Q.Q. Tan, "A deeper look at machine learning-based cryptanalysis," Annual International Conference on the Theory and Applications of Cryptographic Techniques, pp.805–835, Springer, 2021.
[6] J.H. Lee, M. Heo, K.R. Kim, and C.S. Kim, "Single-image depth estimation based on fourier domain analysis," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.330–339, 2018.
[7] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., pp.32–36, IEEE, 2004.
[8] Y. Chen, Y. Shen, H. Yu, and S. Yuan, "A new neural distinguisher considering features derived from multiple ciphertext pairs," The Computer Journal, vol.66, no.6, pp.1419–1433, 2023.
[9] Z. Hou, J. Ren, and S. Chen, "Improve neural distinguishers of simon and speck," Security and Communication Networks, vol.2021, pp.1–11, 2021.
[10] L. Zhang, Z. Wang, J. Guo, *et al.*, "Improving differential-neural cryptanalysis with inception," Cryptology ePrint Archive, 2022.
[11] J. Liu, J. Ren, S. Chen, and M. Li, "Improved neural distinguishers with multi-round and multi-splicing construction," Journal of Information Security and Applications, vol.74, p.103461, 2023.
[12] R. Beaulieu, D. Shors, J. Smith, S. Treatman-Clark, B. Weeks, and L. Wingers, "The simon and speck lightweight block ciphers," Proceedings of the 52nd annual design automation conference, pp.1–6, 2015.
[13] E. Biham and A. Shamir, "Differential cryptanalysis of des-like cryptosystems," Journal of CRYPTOLOGY, vol.4, no.1, pp.3–72, 1991.
[14] K. Fu, M. Wang, Y. Guo, S. Sun, and L. Hu, "Milp-based automatic search algorithms for differential and linear trails for speck," International Conference on Fast Software Encryption, pp.268–288, Springer, 2016.
[15] S. Sun, L. Hu, P. Wang, K. Qiao, X. Ma, and L. Song, "Automatic security evaluation and (related-key) differential characteristic search: application to simon, present, lblock, des and other bit-oriented block ciphers," International Conference on the Theory and Application of Cryptology and Information Security, pp.158–178, Springer, 2014.
[16] C. Blondeau and B. Gérard, "Multiple differential cryptanalysis: theory and practice.," FSE, pp.35–54, Springer, 2011.
[17] T. Yadav and M. Kumar, "Differential-ml distinguisher: Machine learning based generic extension for differential cryptanalysis," International Conference on Cryptology and Information Security in Latin America, pp.191–212, Springer, 2021.
[18] R.M. Aziz, M.F. Baluch, S. Patel, and A.H. Ganie, "Lgbm: a machine learning approach for ethereum fraud detection," International Journal of Information Technology, vol.14, no.7, pp.3321–3331, 2022.
[19] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously.," J. Mach. Learn. Res., vol.20, no.177, pp.1–81, 2019.
[20] A. Gohr, G. Leander, and P. Neumann, "An assessment of differential-neural distinguishers," Cryptology ePrint Archive, 2022.
[21] A. Baksi, "Machine learning-assisted differential distinguishers for lightweight ciphers," in Classical and Physical Security of Symmetric Key Cryptographic Algorithms, pp.141–162, Springer, 2022.

Gao Wang received the B.S. degree in Software Engineering from Shandong University of Science and Technology, in 2019. He is currently pursuing the master's and Ph.D. degree with East China Normal University, Shanghai, China. He research interests include symmetric cryptography, differential attack, and network security.

Gaoli Wang received the B.S. degree in fundamental mathematics and the Ph.D. degree in information security from Shandong University, Jinan, China. She is currently a Professor with the Software Engineering Institute, East China Normal University. Her research interests include cryptography, computer, and network security.



Siwei Sun received the B.S. degree in Beijing University Of Technology and the Ph.D. degree in University of Chinese Academy of Sciences, Beijing, China. He is currently a Professor with the School of Cryptology, University of Chinese Academy of Sciences. His research interests include automation of symmetric cryptographic algorithm design and analysis, optimization and secure implementation of cryptographic algorithms, and symmetric cryptanalysis based on quantum computing.