A Survey of Side-Channel Leakage Assessment

Yaru Wang 1,2 and Ming Tang 1,2,*

1 School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China; 2018102110055@whu.edu.cn
2 Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, Wuhan University, Wuhan 430072, China
* Correspondence: m.tang@126.com; Tel.: +86-153-2741-2612

Abstract: As more threatening side-channel attacks (SCAs) are being proposed, the security of cryptographic products is seriously challenged. This has prompted both academia and industry to evaluate the security of these products. The security assessment is divided into two styles: attacking-style assessment and leakage detection-style assessment. In this paper, we will focus specifically on the leakage detection-style assessment. Firstly, we divide the assessment methods into Test Vector Leakage Assessment (TVLA) and its optimizations and summarize the shortcomings of TVLA. Secondly, we categorize the various optimization schemes for overcoming these shortcomings into three groups: statistical tool optimizations, detection process optimizations, and decision strategy optimizations. We provide concise explanations of the motivations and processes behind each scheme, as well as compare their detection efficiency. Through our work, we conclude that there is no single optimal assessment scheme that can address all shortcomings of TVLA. Finally, we summarize the purposes and conditions of all leakage detection methods and provide a detection strategy for actual leakage detection. Additionally, we discuss the current development trends in leakage detection.

Keywords: leakage assessment technology; side channel attack; TVLA; leakage detection

1. Introduction

The pervasive nature of information technology has permeated all aspects of work and life. As malicious information security incidents like “Eternal Blue” [1] and “Bwp47” [2] continue to emerge, information security has garnered significant attention. The cryptographic products are commonly known as products that utilize cryptography technology. The security is the fundamental attribute of cryptographic products. However, various cryptographic analysis technologies, such as traditional cryptographic analysis and SCAs, can impact the security of these products. The traditional cryptographic analysis technology mainly includes techniques like differential cryptanalysis [3], linear cryptanalysis [4], correlation analysis [5], etc. On the other hand, SCA techniques encompass power analysis attacks [6] (such as simple power analysis (SPA) [7], differential power analysis (DPA) [6], correlation power analysis (CPA) [8], mutual information analysis (MIA) [9], and SCA-based deep learning [10–16]), timing attacks [17], fault-based attacks [18], cache attacks [19], etc. Consequently, evaluating the security of cryptographic products has become a crucial task. Two popular security certification standards, namely, Common Criteria (CC) [20] and FIPS [21], have been established to assess the security of cryptographic products. These security certification standards employ two distinct assessment methods: evaluation-style testing (also known as attacking-style assessment) and conformance-style testing (also known as leakage detection-style assessment) [22].

The attacking-style assessment mainly uses various SCA techniques to obtain the information of keys and evaluate the products against SCAs. The attacking-style assessors can directly obtain the key by executing SCAs. The evaluation results facilitate the calculation of security metrics for encryption algorithms and the identification of vulnerabilities...
within these algorithms and provide guidance for implementing algorithm protection strategies. The attacking-style assessment provides the advantages of rigorous evaluation intensity and easily accessible evaluation results. The attacking-style assessment holds great significance as a method for evaluating side-channel security. However, the use of the attacking-style assessment in primary security assessments is considered unsuitable due to its reliance on SCA technologies, which necessitate evaluators with advanced SCA techniques. This reliance increases both the time and sample complexity of the assessment, subsequently slowing down the assessment speed. Consequently, it fails to keep pace with the rapid cycle of product innovation [23–27]. Instead, the leakage detection-style assessment is proposed as a preliminary method to evaluate the security of cryptographic products. The leakage detection-style assessment primarily relies on information theory or statistical hypotheses to analyze side-channel measurements and determine the presence of any side-channel leakage. The objective of the leakage detection-style assessment is to evaluate whether the device can pass the testing rather than recover the key itself. This type of assessment can be categorized into two main types: leakage assessment based on information theory and assessment based on statistical hypothesis. In recent years, the leakage assessment based on statistical hypotheses has been widely adopted due to its superior efficiency and effectiveness [27]. Various works focusing on leakage assessment using statistical hypotheses have been proposed [23–38].

In this paper, we study the works of side-channel leakage assessment and provide succinct accounts of motivations and detection processes behind these assessment methods. We also compare the efficiency of different detection methods and discuss the future development of leakage assessments.

The main contributions of this paper are as follows:

(1) We analyze the works of side-channel leakage assessment and classify the leakage detection-style assessment works into two categories: the technology of TVLA and optimizations of TVLA. Additionally, we identify the shortcomings of TVLA. Due to the TVLA’s flaws of statistical tool, detection process, and decision strategy, we divided TVLA’s optimization schemes into three groups: the optimization of statistical tool, detection process, and decision strategy. Furthermore, we provide a brief description of the motivation and detection process for each optimization and compare their detection efficiency.

(2) Due to the lack of a unified and comprehensive leakage detection assessment method that can address all the TVLA’s shortcomings, as well as the variation in optimization methods based on detection purposes and conditions, we present a summary on how to select a suitable leakage detection assessment method depending on specific detection purposes and conditions. Moreover, considering the current state of leakage detection assessment, we discuss potential future trends in this field.

The structure of this paper is as follows. Section 2 provides a brief description of the process, methods, metrics, and shortcomings of the attacking-style assessment. Section 3 focuses on the leakage detection-style assessment and describes the development process, metrics, and detection objectives. Section 4 mainly focuses on the leakage detection assessment based on statistical hypotheses and describes TVLA and its optimization methods. Section 5 highlights the relationship between the attacking-style assessment and leakage detection-style assessment. In Section 6, the current status and future development trends of leakage detection-style assessment are discussed. Finally, Section 7 presents the conclusion of this review.

2. The Attacking-Style Assessment

2.1. The Assessment Process of Attacking-Style Assessment

The attacking-style assessment primarily utilizes the device’s side-channel information to recover the key and assess the product’s security against SCAs. The CC [20] standard serves as a reference criterion for the attacking-style assessment. The estimators can choose
any SCA method from the list of threats to conduct SCAs and assess the product’s security level. Figure 1 illustrates the process of the attacking-style assessment.

**Figure 1.** The process of attacking-style assessment.

In the attacking-style assessment, the evaluator assumes the role of an attacker with prior knowledge of the implementation of cryptographic algorithms.

2.2. The Methods of Attacking-Style Assessment

Because the attacking-style assessment heavily relies on SCA technologies, the evaluators must have a proficient understanding of SCA technologies. SCA techniques can be categorized into two groups: profiled attack (PA) and non-profiled attack (NPA). The attackers using PAs must first construct a device model before conducting SCAs. The current methods for PAs include Template Attack (TA) [39–41] and the profiled attack-based deep learning [14,42,43]. On the other hand, NPA does not require a model and solely relies on side-channel measurements for conducting SCAs. Common NPA methods include DPA [7], CPA [8], MIA [9], etc.

2.2.1. The Profiled Attack

(1) The Template Attack

Time series are the typical representation of side-channel measurements. The attackers utilize these measurements along with their model to conduct side-channel power analysis. TAs are one of the earliest approaches to PAs [39–41]. In TAs, it is assumed that the attacker has access to the same device as the target device, enabling them to encrypt any plaintext and collect corresponding side-channel power traces. TAs include two stages: the stage of constructing the template and the stage of matching the template.
In the stage of constructing the template, the main objective is to extract trace features and build the template. Assuming we have \( n \) traces designated as \( L = \{ l_1, l_2, \ldots, l_n \} \), we divide \( L \) into \( K \) groups noted as \( L_1, L_2, \ldots, L_K \), where \( L_i \) is the \( i \)-th group with \( k_i \). The template of \( k_i \) is noted as \( (\mu_i, C_i) \), where \( \mu_i \) represents the mean vector and \( C_i \) is the covariance matrix.

\[
\mu_i = \frac{\sum_{j \in A_i} l_j}{n_i}, \quad C_i = \frac{1}{n_i-1} \sum_{j \in A_i} (l_j - \mu_i)(l_j - \mu_i)^T.
\]  

During the stage of matching the template, the attacker utilizes the traces to calculate the probability between the trace and the templates \( (\mu_i, C_i) \).

\[
P(\text{trace}; (\mu_i, C_i)) = \frac{\exp \left( -0.5 \times (\text{trace} - \mu_i)^T C_i^{-1} (\text{trace} - \mu_i) \right)}{\sqrt{(2\pi)^2 \det(C_i)}}.
\]  

If \( P(\text{trace}; (\mu_i, C_i)) > P(\text{trace}; (\mu_j, C_j)), \forall i \neq j \), then we get \( k_{\text{guess}} = k_j \).

2.2.2. The Non-Profiled Attack

The fundamental assumption of NPAs is that the attacker cannot obtain the same device as the target device but has access to an unlimited number of side-channel traces.

1. Differential Power Analysis

The key information is extracted by the distinguisher in NPA. Let \( X = (X_1, \ldots, X_D)' \) be the input, where \( X_i \) represents the \( i \)-th group. The attacker collects the traces from the process of encrypting \( D \) groups’ data. The trace of \( X_i \) is denoted as \( l_i' = (l_{1,i}, \ldots, l_{n,i}) \), where \( |K| \) is the number of all possible keys. For every key, the input \( X \) and the intermediate value \( V = f(X, k) \) are considered. For input \( X \) and all guessing keys \( K \), the trace matrix \( L \) and intermediate value matrix \( V \) are

\[
L = \begin{bmatrix}
  l_{1,1} & l_{1,2} & \cdots & l_{1,i} & \cdots & l_{1,n} \\
  l_{2,1} & l_{2,2} & \cdots & l_{2,i} & \cdots & l_{2,n} \\
  \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
  l_{i,1} & l_{i,2} & \cdots & l_{i,i} & \cdots & l_{i,n} \\
  l_{D,1} & l_{D,2} & \cdots & l_{D,i} & \cdots & l_{D,n}
\end{bmatrix}, \quad V = \begin{bmatrix}
  v_{1,1} & v_{1,2} & \cdots & v_{1,i} & \cdots & v_{1,n} \\
  v_{2,1} & v_{2,2} & \cdots & v_{2,i} & \cdots & v_{2,n} \\
  \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
  v_{i,1} & v_{i,2} & \cdots & v_{i,i} & \cdots & v_{i,n} \\
  v_{D,1} & v_{D,2} & \cdots & v_{D,i} & \cdots & v_{D,n}
\end{bmatrix}
\]
Map the intermediate value matrix $V$ to matrix $H$ and calculate the correlation coefficient between $h_i$ of $H$ and $I_j$ of $L$. The correlation coefficients $r_{ij}$ are stored in matrix $R$.

$$
H = \begin{bmatrix}
    h_{1,1} & h_{1,2} & \ldots & h_{1,j} & h_{1,K} \\
    h_{2,1} & h_{2,2} & \ldots & h_{2,j} & h_{2,K} \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
    h_{D,1} & h_{D,2} & \ldots & h_{D,j} & h_{D,K}
\end{bmatrix},
R = \begin{bmatrix}
    r_{1,1} & r_{1,2} & \ldots & r_{1,j} & r_{1,n_i} \\
    r_{2,1} & r_{2,2} & \ldots & r_{2,j} & r_{2,n_i} \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
    r_{k,1} & r_{k,2} & \ldots & r_{k,j} & r_{k,n_i}
\end{bmatrix},
$$

$$
r_{ij} = \frac{\sum_{d=1}^{D}(h_{i,d} - \bar{h}_i)(l_{j,d} - \bar{l}_j)}{\sqrt{\sum_{d=1}^{D}(h_{i,d} - \bar{h}_i)^2} \cdot \sum_{d=1}^{D}(l_{j,d} - \bar{l}_j)^2}.
$$

We use correlation coefficients to assess the linear correlation between $I_j$ and $h_i$, where $i = 1, \ldots, K$, $j = 1, \ldots, n_i$ in DPA. As $r_{ij}$ increases, the level of correspondence between $I_j$ and $h_i$ also augments. By identifying the highest value of $r_{ij}$, the assailant can successfully retrieve the key.

(2) Correlation Power Analysis

Correlation Power Analysis (CPA) is a variant of DPA that exploits the correlation between the power traces $L$ and the leakage model $F$ to conduct the attack. Assuming the leakage model is denoted by $F$, we map the intermediate value matrix $V = f(X, K)$ to the leakage value noted as $G_K = F(V)$. Here, $X$ represents the input, $K$ denotes the key space, and $f$ stands for the cryptographic function. Typically, $F$ represents the leakage model based on either the Hamming weight or Hamming distance. Similarly, $G_K$ corresponds to the Hamming weights or Hamming distances of $V$.

The correlation coefficient between the power traces $L$ and $G_K$ is as follows:

$$
p_{L,G_K} = \frac{E(L \cdot G_K) - E(L) \cdot E(G_K)}{\sigma_L \cdot \sigma_{G_K}},
$$

where $E(L)$, $E(G_K)$, $E(L \cdot G_K)$ are the expectations of $L$, $G_K$, $L \cdot G_K$ and $\sigma_L$, $\sigma_{G_K}$ are the variances of $L$, $G_K$, respectively.

The attackers iterate through the space of possible keys, calculating the correlation coefficient between $G_K$ and $L$ to determine whether a key guess is correct. The correlation coefficient of the correct guess key is higher than that of the incorrect guess key. The attackers identify the guess key with the maximum correlation coefficient as the correct key, using the maximum likelihood estimation method.

(3) Mutual Information Analysis

The attacker of MIA uses mutual information or entropy to assess the correlation between the leakage and the intermediate value or between the intermediate value and side-channel measurements in [9]. The fundamental concepts of entropy and mutual information are as follows:

Let $X = (X_1, X_2, \ldots, X_n)$ be the set of discrete random variables, then the entropy of $X$ is

$$
H(X) = \sum_{x \in X} -p(x) \log p(x)
$$

where $p(x)$ is the probability distribution of variables.

Let $X = (X_1, X_2, \ldots, X_n)$, $Y = (Y_1, Y_2, \ldots, Y_n)$ be the set of discrete random variables, then the conditional entropy $H(X|Y)$ of $X$ under $Y$ is

$$
H(X|Y) = -\sum_{y \in Y} p(y) \log p(y) \sum_{x \in X} p(x|y) \log p(x|y).
$$

The mutual information between $X$ and $Y$ is $I(X; Y) = H(X) - H(X|Y)$. 


Let $K$, $X$, and $V$ be the key, plaintext, and intermediate value, respectively. The correct key is denoted as $k_c$, and the leakage function $L(V) = f(X, K)$ is continuous. Let $l_i = f(X_i, K_c)$, where $1 \leq i \leq N$. For a given key $k$, the intermediate values can be obtained using $M = (X_i, k)$, where $M$ represents discrete mutual information. When the attack is successful, $\max_{k \in K}(|D(M(p, k), L)) = k_c$, $D$ is the distinguisher.

In MIA, the distinguisher was proposed as

$$I(L, M(X, k)) = H(I) - H(I|M(X, k)).$$

(7)

Different keys result in different values of mutual information. The average mutual information is associated with the conditional entropy $H(I|M(X, k))$. A stronger correlation between the measured $L$ and $M(X, k)$ indicates that the guessed key $k_{\text{guess}}$ is the correct key.

The development of the attacking-style assessment relies entirely on SCA methods. Essentially, the attacking-style assessment comprises attacks on the targeted devices. Consequently, the assessment itself is considered as an attack.

2.3. The Metrics of Attacking-Style Assessment

Due to the assessment evaluation being an attack, the attack metrics are the assessment metrics for attacking-style assessment. Distinguisher scores are commonly employed to sort the candidate key $k_{\text{guess}}$ in an SCA. The position of the correct key $k_c$ in the sorting results is called the key ranking, noted as $\text{rank}(k_c | L, X)$. The metrics based on the key ranking are defined as follows.

1. The number of samples: Find a minimum positive integer $N$, when the sample size $|L| \geq N$, and there is

$$\text{rank}(k_c | L, X) \leq |\hat{K}_c|.$$  

(8)

where $|\hat{K}_c|$ is key ranking of $k_c$. Especially, when $|\hat{K}_c| = 1$, $k_c = \hat{K}_c$, then $\text{rank}(k_c | L, X) = 1$.

2. Success rate: The success rate of the side-channel attack is the probability of successfully recovering the correct key.

3. Guessing entropy: The guessing entropy is the mathematical expectation of key ranking

$$GE_{L,X} = \mathbb{E}(\text{rank}(K_c | L, X)).$$

(9)

2.4. The Advantages and Shortcomings of Attacking-Style Assessment

With the advancement of side-channel technology, there has been a rise in the proposal of various side-channel attack methods. Evaluators can select an appropriate side-channel attack method to evaluate cryptographic algorithms, considering distinct encryption implementations and attack conditions. Through evaluating the actual results of side-channel attacks, one can determine the security level of the encryption implementation. The attacking-style assessment technique is characterized by its high level of aggression and extensive evaluation, allowing evaluators to directly acquire information about key and design weaknesses. The attacking-style assessment enables a thorough comprehension of system vulnerabilities and weaknesses. Through simulating real-world attack scenarios, it offers valuable insights into the efficacy of security measures and aids in identifying potential areas for improvement. It facilitates the identification of previously unknown vulnerabilities. The capability to analyze attacks in a controlled environment enables the development and implementation of efficient countermeasures. Thus, the attacking-style assessment is better suited for conducting comprehensive analysis and evaluation of cryptographic algorithms, subsequently facilitating vulnerability analysis and the establishment of protective measures.
The attacking-style assessment is widely utilized in the security assessment of cryptographic products. However, there are several limitations to its applicability. First, because new SCA methods are frequently proposed, it is crucial to periodically update the list of attack methods. However, CC standards generally apply to high-security products like bank smart cards and passport ICs. The time and computational complexity involved in security evaluation make it difficult to keep pace with the innovation cycle of new security products [26,45,46]. Second, the evaluators of attacking-style assessments must possess exceptional expertise in SCA methods and measurement techniques [23]. Third, due to the different principles of SCAs, evaluators need to perform multiple SCAs to calculate security metrics. The increasing number of attacks inevitably leads to an increase in computational and time complexity [31,39]. Therefore, the attacking-style assessment is not suitable for the primary security evaluation of cryptographic algorithms. Instead, the leakage detection-style assessment is proposed as a preliminary method to evaluate the security of cryptographic products.

### 3. The Leakage Detection-Style Assessment

#### 3.1. The Goals of Leakage Detection-Style Assessment

The leakage detection-style assessment is conducted by the laboratory to provide the security certification, or conducted by the designers during the design period to highlight and address the potential issues before the product is launched into the market. Consequently, different stages of assessment have different goals. There are four different intentions [47]: certifying vulnerability, certifying security, demonstrating an attack, and highlighting vulnerabilities.

1. **Certifying vulnerability**: This involves identifying at least a leakage point in traces. It is crucial to minimize the false positive rate.
2. **Certifying security**: The goal here is to find no leakages after thorough testing. In this case, the false negatives become a concern. It is important to design the tests with “statistical power” to ensure a large enough sample size for detecting effects with reasonable probability. Moreover, all possible intermediates and relevant higher-order combinations of points should be targeted.
3. **Demonstrating an attack**: The objective is to map a leaking point (or tuple) to its associated intermediate state(s) and perform an attack. The reporting of the outcomes or projections derived from these attacks is of interest. The false positives are undesirable as they represent wasted efforts of the attacker.
4. **Highlighting vulnerabilities**: The purpose is to map all exploitable leakage points to intermediate states in order to guide designers in securing the device. This has similarities to certifying security, as both require exhaustive analysis. The false negatives are of greater concern than false positives, as false negatives indicate unfixed side-channel vulnerabilities.

#### 3.2. The Process of Leakage Detection-Style Assessment

In leakage detection, the evaluators have the ability to control both input and output variables and obtain the side-channel measurements. The process of leakage detection-style assessment primarily encompasses four stages: collecting power traces, categorizing power traces, calculating the statistical moment, and determining leakage. The process of leakage detection-style assessment is shown in Figure 2.
becomes an ideal method for primary security assessments due to the lowered technical threshold and reduced evaluation time.

3.3. The Development of Leakage Detection-Style Assessment

In this paper, we categorize the development of leakage detection-style assessment into three stages: the phase of proposing the leakage assessment concept, the phase of forming leakage assessment methods, and the phase of optimizing the leakage assessment methods. Figure 3 provides a visual representation of these development stages.

In the phase of proposing the leakage assessment concept, Corona (2001) proposed the idea of security assessment [48]. The failed result indicates the presence of side-channel leakage, whereas the pass result does not imply that there is no leakage but rather indicates that the leakage has not been identified at the specified confidence level $\alpha$. In [49], a framework for the side-channel security assessment was presented, categorizing security assessments into attacking-style and leakage detection-style assessments. This paper primarily concentrates on the leakage detection-style assessment.

In the phase of forming leakage assessment methods, two groups of leakage detection-style assessment methods have emerged based on different theories: the information theory and the statistical hypothesis. In 2010 and 2011, Chothia proposed two leakage assessment methods: one based on discrete mutual information (DMI) [50] and the other based on continuous mutual information (CMI) [51]. Both the DMI and CMI methods utilize information entropy as a testing tool to assess the possibility of side-channel leakage. Additionally, Gilbert et al. utilized the statistical hypothesis as a testing tool in their work [52]. They divided the traces into two groups and performed leakage assessment on the Advanced Encryption Standard (AES) using a $t$-test [52]. In 2013, based on the research by Gilbert et al. [52], Becker et al. proposed the Test Vector Leakage Assessment (TVLA) technique [53]. They divided the traces into two groups: fixed plaintext traces vs. random plaintext traces. Then, they performed Welch’s $t$-tests on these two groups to detect any mean difference between the trace sets of fixed plaintext and random plaintext. In 2013, Mather et al. conducted a comparison of the detection efficiency between DMI and CMI with TVLA [27]. The results demonstrated that TVLA outperformed DMI and
CMI in terms of detection efficiency. Consequently, the leakage assessment based on the statistical hypothesis has been developed in recent years. In 2015, Moradi et al. summarized TVLA techniques proposed by Gilbert and Becker in [23] and studied leakage detection in various scenarios. The non-specific TVLA technology has been widely applied in leakage assessment and is commonly regarded as a preliminary assessment technique in both industry and academia due to its simplicity, efficiency, and versatility.

Figure 2. The process of leakage detection-style assessment.

However, in order to address the shortcomings of TVLA [22–40,47] (the detailed description of TVLA's drawbacks can be found in Section 4), improve the detection efficiency, reliability of results, and quantify side-channel vulnerability, the researchers have proposed a variety of leakage assessment schemes to enhance the TVLA.

In the phase of optimizing leakage assessment methods, there are three main aspects to consider. Firstly, the optimization of statistical tools aims to replace the $t$-test of TVLA with alternative statistical tools (such as the paired $t$-test [26], $\chi^2$-tests [28], Hotelling $T^2$-tests [29], KS tests [30], ANOVA [31], deep learning [36,37], etc.). Secondly, the optimization of the detection process involves improving the current TVLA detection process or proposing a new one [32–34]. Finally, the optimization of the decision strategy focuses on introducing a new decision strategy, such as the HD strategy [35], to determine the stage of leakage.

4. The Leakage Assessment Based on Statistical Hypothesis

This section provides an overview of two perspectives on leakage assessment technologies: the TVLA technology and the TVLA’s optimization schemes.

4.1. The Test Vector Leakage Assessment

In 2013, the Cybersecurity Research Institute (CRI) introduced the TVLA as a standardized approach for detecting side-channel leakages. This section emphasizes the process of detection, the metrics used for detection, and discusses the limitations of TVLA.
4.1.1. The TVLA Technology

(1) The detection process of TVLA

The detection process of TVLA is as follows.

In the stage of collecting power traces, \( N \) different inputs \( X \) are used to collect traces \( L \) from the execution process. Let \( l = \{l_1, \ldots, l_i, \ldots, l_{n_l}\} \) be a trace, where \( l_i \) represents the measurements of the \( i \)-th point, and \( n_l \) is the length of traces.

In the stage of categorizing power traces, the trace set \( L \) is divided into two groups: fixed plaintext trace set \( L_A \) and random plaintext trace set \( L_B \). It is assumed that \( L_A \) and \( L_B \) obey the normal distribution \( \mathcal{N}(\mu_A, \sigma_A^2) \) and \( \mathcal{N}(\mu_B, \sigma_B^2) \), respectively. The cardinality, sample mean, and sample variance of \( \mu_A \) are denoted as \( (n_A, \bar{X}_A, \sigma_A^2) \), while those of \( L_B \) are denoted as \( (n_B, \bar{X}_B, \sigma_B^2) \).

In the stage of calculating the statistical moment, the null hypothesis \( H_0 \) states that there is no side-channel leakage, while the alternative hypothesis \( H_1 \) suggests the presence of leakage. Welch’s \( t \)-test is employed to determine the mean difference between \( L_A \) and \( L_B \). Based on the assumption of \( H_0 \), we calculate the statistical moment \( T_l \) and probability \( P_l \) of accepting \( H_0 \).

\[
P_l = P(|T_l| > |T_{th}|). \tag{10}
\]

In the stage of determining leakage, if \( |T_l| > |T_{th}| \), then the null hypothesis \( H_0 \) is rejected, and it can be concluded that there is side-channel leakage.

(2) The statistical tool of TVLA

The statistical tool Welch’s \( t \)-test is used to test the mean difference between \( L_A \) and \( L_B \) in TVLA. The null hypothesis \( H_0 \) and alternative hypothesis \( H_1 \) in Welch’s \( t \)-test [33] are as follows:

\[
H_0 : \mu_A = \mu_B, \quad H_1 : \mu_A \neq \mu_B
\]

The statistical moment \( |T_{th}| \) and freedom degree \( v \) of the \( t \)-test are calculated as follows.

\[
|T_{th}| = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B} v}}
\]

\[
|T_{th}| = \frac{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} v}{\frac{s_A^2}{n_A - 1} + \frac{s_B^2}{n_B - 1}}. \tag{11}
\]

The probability \( p \) value of accepting hypothesis \( H_0 \) is calculated as follows using the probability density function (PDF).

\[
f(t, v) = \frac{\Gamma\left(\frac{v + 1}{2}\right)}{\sqrt{\pi v} \Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \quad P = 2 \int_{|T_{th}|}^{\infty} f(t, v) dt, \tag{12}
\]

where \( \Gamma(x) = \int_0^\infty t^{x-1}e^{-t} dt \) is the gamma function.

The threshold 4.5 [54,55] is employed to determine the acceptance of \( H_0 \). If \( |T_{th}| > 4.5 \), the hypothesis \( H_0 \) is rejected. When \( v > 1 \ 000, \ P = 2 \times f(4.5, v) < 10^{-5} \) [33], and this implies that the probability of accepting hypothesis \( H_0 \) is less than 0.00001, while the probability of rejecting hypothesis \( H_0 \) is greater than 0.999999, indicating the presence of side-channel leakage in the device.

(3) The decision strategy of TVLA

Due to TVLA being a leakage assessment method for univariate data, when the evaluator obtains the traces of length \( n_L \), TVLA is applied to each sample point of the traces. The assessor can obtain \( n_L \) detection results, and the assessment decision is obtained by combing these results. The \( \min - P \) strategy is a common decision strategy used in TVLA. Let \( L = \{l_1, \ldots, l_{n_L}\} \) be set of traces, where \( l_i = \{l_{i,1}, \ldots, l_{i,j}, \ldots, l_{i,N_i}\} \). The \( T_{th} \) represents the statistical moment value of \( l_i \) at \( i \).
When TVLA is applied to the long traces, the assessor actually conducts multiple ($n_L$ times) tests. If any of the tests reject $H_0$, it indicates the presence of side-channel leakage. In other words, the statistical moment is compatible with $\max_{1 \leq i \leq n_L} |T_{\mu_i}| \geq th$, or the minimum $p$-value is less than the threshold $\alpha_{th}$. This indicates a leakage, meaning that the $\min - P$ strategy of TVLA only uses one test result (the minimum $p$-value) to make a decision regarding long traces.

4.1.2. The Assessment Metrics of TVLA

Due to the fact that TVLA involves the statistical analysis of random variables, it is possible for the detection results to contain errors. Therefore, it becomes imperative to evaluate the detection results. Commonly used metrics are employed to assess the effectiveness of detection methods.

(1) The number of samples: The minimum sample size required to exceed the threshold for statistical moments is an assessment metric used in TVLA. This metric is frequently utilized to compare the detection effectiveness among various assessment methods. In identical conditions, a smaller sample size indicates a higher degree of assessment effectiveness [24–31].

(2) The false positive and false negative: The two types of errors commonly encountered in hypothesis testing are false positives [24] and false negatives [47]. A false positive occurs when the null hypothesis is true, but it is rejected by a $t$-test, leading to an incorrect conclusion. In the context of leakage detection, a false positive refers to a situation where the device does not have any leaks, yet the TVLA results indicate otherwise. Conversely, a false negative denotes a Type II error, which occurs when the null hypothesis is false, but the $t$-test fails to reject it. The rate of Type II errors is denoted as $\beta$. During leakage assessment, the assessor aims to control the false positive rate at the specified significance level $\alpha$.

(3) The effect size: The effect size $\zeta$ is an indicator [22,47] employed to assess the effectiveness of leakage detection. Cohen’s $d$ [56,57] is a commonly used effect size for comparing differences among groups, mainly applied to $t$-tests to compare the standard difference between two means. Cohen’s $d$ is computed as follows:

$$d = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A+n_B-2}}}$$

where $\bar{x}_A$, $\bar{x}_B$ are the sample mean, $s_A^2$, $s_B^2$ are the sample variance, and $n_A, n_B$ are the cardinalities. Cohen establishes the threshold as the criterion for judging the effect size [58]. According to his classification, when $d \leq 0.2$, the effect size is considered “small”, and when $d \leq 0.8$, the effect size is considered “large”.

(4) The power: The power is an essential metric that indicates the ability of the assessor to detect a difference at the significance level $\alpha$, and it is noted as $1 - \beta$. The power should not be less than 75% and is typically required to reach 80% or 90%. The relationship among the variance $\sigma_1^2$, $\sigma_2^2$, the power $1 - \beta$, the significance level $\alpha$, the effect size $\zeta$, and the number of samples $N$ is as follows:

$$N = 2 \left( \frac{T_{\alpha/2} + T_\beta}{\zeta^2} \right)^2 \left( \sigma_1^2 + \sigma_2^2 \right)$$

where $\zeta = \mu_1 - \mu_2$, and $T_{\alpha/2}$ and $T_\beta$ are the statistical moments. Equation (14) allows us to obtain any one of the significance level, effect size, or power [47].

(5) The observed power: Because power is a theoretical parameter that cannot be directly obtained from samples, in order to evaluate the reliability of assessment results, the observed power (OP) is proposed in [59] for given $\alpha$ and $N$. The OP can be considered an approximation of power. If the probability of accepting the null hypothesis $P > \alpha$, the
assessor can determine the reliability of assessment results by checking whether OP is greater than 0.8. Additionally, OP can serve as a means to compare the effectiveness of different assessment methods, while keeping the sample size \( N \) constant.

4.1.3. The Drawbacks of TVLA

Generally, TVLA is based on hypothesis testing to determine whether the side-channel measurements expose any secret information, which can help simplify the security assessment process [25,60–62]. However, this complexity reduction is accompanied by an increase in false positives or false negatives [23,36]. The drawbacks of TVLA are as follows:

(1) Difficulty interpreting negative outcomes

Formally, a statistical hypothesis test can either reject the null hypothesis or fail to reject it. However, it cannot prove or accept the null hypothesis. If the statistical hypothesis test fails to reject the null hypothesis, the security assessment agency must demonstrate that their assessment method is fair. Unfortunately, due to the sample limitations or time constraints, varying levels of expertise, and poor equipment quality, the fairness of leakage assessment may be undermined. As a result, when the negative outcomes occur, it becomes challenging to explain and provide evidence for the fairness of TVLA.

(2) Unreliability of positive outcomes

The TVLA is commonly utilized for univariate analysis. However, in actual detection scenarios, multiple tests are necessary. To illustrate this, let us assume that the probability of false positive is denoted as \( \alpha \). The length of traces is represented as \( n_l \), and there are \( n_l \) independent tests. Consequently, the probability of rejecting the null hypothesis for a single test is given by \( \alpha_{all} = 1 - (1 - \alpha)^n_l \). Ding has emphasized that the threshold of 4.5 for TVLA corresponds to a false positive rate of approximately \( \alpha \approx 0.00001 \). For instance, if there are 1000 independent tests, then \( \alpha_{all} = 0.0068 \). On the other hand, if there are 1,000,000 independent tests, then \( \alpha_{all} = 0.9987 \) in [63]. Therefore, the products that generate long power traces are more likely to be deemed vulnerable compared to those with shorter traces. Consequently, the positive outcomes may be unreliable.

(3) Impossibility of achieving exhaustive coverage

Ideally, an evaluator would prefer to eliminate any possible sensitive dependency across all distributional forms, for all points and tuples of points as a whole, considering all target functions and intermediate states before determining the security of a target device. However, even when the best efforts are made, there are still limitations. Moreover, in order to enhance the scope of detection, extensive tests are required, which lead to an increase in the type I error rate, computational complexity, and sample complexity. Consequently, achieving exhaustive coverage through TVLA is not possible.

(4) The multivariate problems of TVLA

In TVLA, it is assumed that the sample points in a trace are independent of each other. Additionally, TVLA is a test for the univariate analysis. However, in reality, numerous examples contradict this assumption. This is especially evident in the case of protective algorithm implementations, where the leakage is mostly observed to be multivariable or horizontal [29,37,47]. Therefore, it is crucial to consider the correlation between multivariate variables in leakage assessment.

(5) The fewer trace groups and dependence of statistical moment

The simplicity and efficiency of TVLA depend on a reduced number of groups, the fixed VS random traces set, and the mean statistical moment [64]. However, in cases where the leakage does not manifest in the mean statistical moment [32] and there is a presence of mean differences among multiple groups, this introduces the risk of both false positives and false negatives [29–33].
(6) The drawbacks of distribution assumption

The assumption is that the power traces obey a normal distribution in TVLA, while in reality, many examples that contradict this assumption can be observed. Especially for the protective algorithm implementations, it is generally necessary to use the combining functions to preprocess the side-channel traces in TVLA. Additionally, the distribution of samples does not obey the normal distribution after preprocessing [59].

(7) The shortcomings of certifying vulnerability

The assessor of TVLA can only answer whether there is a side-channel leakage but cannot provide information regarding the specific location of the leakage or how to exploit it for key recovery. The results obtained from TVLA are insufficient for certifying vulnerability or deducing the relationship between the detected leakage and attack [22].

4.2. The Optimizations of TVLA

In order to address the shortcomings of TVLA mentioned above, researchers have proposed various optimization assessment methods. This section provides a summary of optimization methods in three aspects, the optimization of statistical tools, the optimization of assessment processes, and the optimization of decision strategies.

4.2.1. The Optimization of the Statistical Tool

The statistical tools play a crucial role in leakage assessments as they significantly impact the detection results. The researchers have attempted to enhance the detection efficiency and reliability of results in TVLA by utilizing alternative statistical tools instead of Welch’s *t*-test when calculating statistical moments. This section summarizes the optimization methods for statistical tools.

(1) The paired *t*-test

The motivation: In [26], Adam Ding found that the environmental noise can adversely affect the results of *t*-tests in actual assessments. In the worst-case scenario, a device with leaks could pass the test solely due to the environmental noise being strong enough to mask them. In order to mitigate the impact of environmental noise, Adam Ding proposed a side-channel leakage detection based on paired *t*-tests in [26], where Welch’s *t*-test was replaced with the paired *t*-tests to eliminate the influence of environmental noise on the results.

The method: In the stage of collecting power traces, a fixed input sequence is proposed to effectively minimize environmental noise. This sequence consists of the repetition of ABBA, such as ABBA ABBA . . . ABBA ABBA. The power traces with minimal environmental variation are obtained through a paired *t*-test.

In the stage of categorizing power traces, the power traces are divided into two sets and noted as $L_A = \{l_A,1, \ldots, l_A, n_A \}$ and $L_B = \{l_B,1, \ldots, l_B, n_B \}$, where $l_{A,i}$ represents the *i*-th trace of $L_A$.

In the stage of calculating the statistical moment, assuming $n_A = n_B = n$, then there are *n* pairs of traces $(l_{A,i}, l_{B,i})$. Let $D = (L_A - L_B) = \{D_1, \ldots, D_n \}$, where $\overline{D}$ and $s_D^2$ represent the sample mean and sample variance of $D$, respectively. Considering the null hypothesis $H_0 : \mu_D = 0$, then the statistical moment $|T_p|$ of the paired *t*-test is as follows:

$$|T_p| = \frac{\overline{D}}{\sqrt{\frac{s_D^2}{n}}}.$$  \hfill (15)

In the stage of determining leakage, use the same threshold and decision strategy as TVLA.

(2) $\chi^2$-test

The motivation: In TVLA, comparing two groups and using the simple mean statistical moment can increase the risk of false negatives [28, 49] or fail to detect leakages [40], when
the leakage does not occur at the mean statistical moment. In 2018, Moradi proposed the side-channel leakage detection based on the $\chi^2$-test [28]. In the $\chi^2$-test, Welch’s $t$-test was replaced with the $\chi^2$-test to detect whether multiple trace groups originate from the same population. Furthermore, in the $\chi^2$-test, the frequency of side-channel measurements is stored in histograms, by analyzing the histograms to identify any distribution differences among the trace groups.

The method: In the stages of collecting power traces and determining leakage, the same method as TVLA is adopted. During the stage of categorizing power traces, the traces are divided into $r$ type groups and each type contains $c$ sets in the $\chi^2$-test. The frequency of side-channel measurements is stored in $c$ and a contingency table is formed. In this table, the $F_{i,j}$ represents the frequency of the $i$-th type group and the $j$-th set. The number of samples is denoted as $N = \sum_{i=0}^{r-1} \sum_{j=0}^{c-1} F_{i,j}$, and $E_{i,j} = \frac{\sum_{i=0}^{r-1} \sum_{j=0}^{c-1} F_{i,j}}{N}$. In the stage of calculating the statistical moment, the null hypothesis $H_0$ of the $\chi^2$-test states that all power traces come from the same population. The statistical moment $T_{\chi^2}$ and the degree of freedom $v$ are obtained by (16).

$$T_{\chi^2} = \sum_{i=0}^{r-1} \sum_{j=0}^{c-1} \frac{(F_{i,j} - E_{i,j})^2}{E_{i,j}}, \quad v = (r - 1)(c - 1),$$

The probability $P$ of accepting $H_0$ is obtained by (17).

$$f(t, v) = \frac{t^{v-1}e^{-t/2}}{2^{v/2}\Gamma(v/2)}, P = 2\int_{|t|}^\infty f(t, v)dt,$$

where $\Gamma(\cdot)$ is the gamma function.

In the stage of determining leakage, use the same threshold and decision strategy as TVLA.

(3) KS test

The motivation: When the leakage does not occur at the mean statistical moment, TVLA is not the optimal choice. Consequently, Zhou X proposed the side-channel leakage detection based on the Kolmogorov–Smirnov (KS) test in [30]. The KS test is a non-parametric statistical test used to determine if two groups of traces originate from the same population by quantifying the distance of cumulative distribution functions between the two groups.

The method: In the stages of collecting power traces, categorizing power traces, and determining leakage, adopt the same method as TVLA. In the stage of calculating the statistical moment, the null hypothesis $H_0$ of the KS test assumes that $L_A$ and $L_B$ come from the same population. The alternative hypothesis $H_1$ states that $L_A$ and $L_B$ come from different populations. The probability $P$ of accepting $H_0$ is as follows.

$$P = 2\sum_{j=1}^\infty (-1)^{j-1}e^{-2j^2Z^2},$$

where $Z = D_{n_A, n_B} \left( \sqrt{1 + \frac{0.11}{\sqrt{J} + 0.12}} \right)$, $J = \frac{n_A n_B}{n_A + n_B}, D_{n_A, n_B}$ represents the maximum distance of cumulative distribution probabilities between $L_A$ and $L_B$, $D_{n_A, n_B} = \sup_x [L_{A, n_A}(x) - L_{B, n_B}(x)]$.

(4) Hotelling $T^2$-test

The motivation: TVLA is based on the assumptions of a sample and its detection efficiency strongly depends on parameters such as the signal-to-noise ratio (SNR), degree of dependency, and density. The correct interpretation of leakage detection results requires prior knowledge of these parameters. However, the evaluators often do not have this prior knowledge, which poses a non-trivial challenge. In order to address this issue, Bronchain.
where \( F \) is the probability of accepting hypothesis \( H_0 \) of the Hotelling \( T^2 \)-test is as follows.

\[
H_0 : \mu_A = \mu_B, \quad H_1 : \mu_A \neq \mu_B
\]

(19)

then, the statistical moment \( T^2 \) is

\[
T^2 = \frac{n_An_B}{n_A + n_B} (\bar{x}_A - \bar{x}_B)^T S^{-1} (\bar{x}_A - \bar{x}_B),
\]

(20)

where \( n_A, n_B \) are the cardinality of trace sets \( L_A \) and \( L_B \), the length of traces is \( n_i \), and the covariance matrix \( C \) is

\[
C = \frac{\sum_{i=1}^{n_A} (x_i - \bar{x}_A)(x_i - \bar{x}_A)^T + \sum_{j=1}^{n_B} (x_j - \bar{x}_B)(x_j - \bar{x}_B)^T}{(n_A + n_B - 2)}.
\]

(21)

The statistical moment \( T^2 \) follows the Fisher distribution with degrees of freedom \((n_1, n_A + n_B - 2)\).

\[
\frac{(n_A + n_B - 1 - n_1)}{(n_A + n_B - 2)n_1} \chi^2_{H_0} = \lambda \chi^2_{n_1} \sim F(n_1, n_A + n_B - 2),
\]

(22)

\[
\lambda = \frac{(n_A + n_B - 1 - n_1)}{(n_A + n_B - 2)n_1},
\]

(23)

then, the probability of accepting hypothesis \( H_0 \) is

\[
P = 1 - F_F (\lambda T^2)
\]

(24)

where \( F_F (x, v_1, v_2) = I_{v_2} (\frac{v_2}{v_1 + v_2}) \), \( I \) is the regularization incomplete \( \beta \) function, and \( v_1 \) and \( v_2 \) are the degrees of freedom.

In the stage of determining leakage, use the same threshold and decision strategy as TVLA.

(5) ANOVA

The motivation: The TVLA and the \( \chi^2 \)-test require a large number of traces to distinguish the difference between leakage points and non-leakage points, and the paired \( t \)-test requires careful selection of inputs and low-noise measurements. Wei Yang proposed a novel leakage detection method using analysis of variance (ANOVA) in [31]. In ANOVA, the traces are categorized into multiple groups, and variance analysis is employed to identify the differences among these groups.

The method: The method of collecting power traces and determining leakage is adopted from TVLA. In the stage of categorizing power traces, the power trace set \( L \) is divided into \( r \) groups. Let \( N, n_i \) be the cardinality of \( L, L_i \), while \( \bar{x} \) represents the sample mean of \( L \) and \( \bar{x}_i \) represents the sample mean of \( L_i \). In the stage of calculating the statistical moment, the null hypothesis \( H_0 \) of the ANOVA test assumes that all group traces come from the same population, then the statistical moment \( T_F \) of ANOVA is calculated as follows.

\[
T_F = \frac{(N - r)S_{Sb}}{(r - 1)S_{Sw}}, \quad S_{Sb} = \sum_{i=1}^{r} \left( \bar{x}_i - \frac{\sum_{i=1}^{n_i} n_i \bar{x}_i}{N} \right)^2, \quad S_{Sw} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2,
\]

(25)
\[ th = F_{1-a}(r-1, N-r), \]  
(26)

The probability \( P \) of accepting hypothesis \( H_0 \) is presented in Equation (27).

\[
P = \int_{|T_f|} f(x,v)dx, \quad f(v_1,v_2,x) = \frac{\Gamma\left(\frac{v_1+v_2}{2}\right)}{\Gamma\left(\frac{v_1}{2}\right) \Gamma\left(\frac{v_2}{2}\right) \left(1 + \frac{v_1}{v_2}\right)^{v_1+v_2}} x^{v_1-1},
\]
(27)

where \( v_1 = r - 1 \) and \( v_2 = N - r \).

6. The deep learning leakage assessment

The motivation: TVLA is conducted under the assumption that each sample point is independent and that the leakage occurs at the mean statistical moment. However, in reality, many examples contradict this hypothesis. Additionally, unaligned traces can also impact the detection results. Therefore, TVLA is inadequate for addressing horizontal and multivariate leakage, as well as unaligned traces. In order to solve this issue, Moos T proposed the method of Deep Learning Leakage Assessment (DL-LA) [37].

The method: DL-LA maintains the basic idea of TVLA, which involves discriminating between two groups of traces, and enhances the side-channel leakage assessment by training a neural network as a distinguisher to discriminate the two group traces. In the stage of collecting power traces, the same method of trace collection as TVLA is implemented. In the stage of categorizing power traces, the trace set \( L \) is divided into the training set and validation set, which do not intersect. The mean \( \mu \) and standard deviation \( \delta \) of the training set are calculated and \( X_j^i = \frac{(X_j - \mu)}{\delta} \) is used to standardize both the training set and validation set, where \( j \) represents the trace and \( i \) represents the sample points of trace. In the stage of calculating the statistical moment, the assessor begins by training the neural network distinguisher, using the training set, and then validating its accuracy using the validation set. If the accuracy of the neural network distinguisher exceeds that of a random guess distinguisher, it can be employed that the neural network distinguisher can be used to discriminate the two groups of traces. The construction of the neural network distinguisher is as follows.

The network is built using Python’s Keras library, with Tensor Flow serving as the backend. It comprises four fully connected layers, consisting of neurons with outputs of 120, 90, 50, and 2, respectively. The ReLU function is utilized as the activation function for the input layer and each inner layer, while softmax functions as the activation function for the final layer. The four layers are separated by a Batch Normalization layer. Once the neural network distinguisher is constructed, the assessor proceeds to employ it to conduct the leakage assessment. The null hypothesis \( H_0 \) states that the traces can be randomly divided into two groups by the neural network distinguisher, and the total number of correct classifications follows the binomial distribution \( X \sim Binom(X, 0.5) \). The probability that the total number of correct classifications is at least \( S_X \) in the pure random distinguisher is \( P(X \geq S_X) \), and the probability is calculated as follows.

\[
P(X \geq S_X) = \sum_{q=S_X}^{X} \binom{X}{q} 0.5^q 0.5^{X-q} = 0.5^X \sum_{q=S_X}^{X} \binom{X}{q}.
\]
(28)

In the stage of determining leakage, the threshold \( P_{th} = 10^{-5} \) is set, and if \( P(X \geq S_X) > P_{th} \), then the hypothesis \( H_0 \) is rejected, indicating the presence of side-channel leakage.

In summary, the various optimization methods (in Table 1) mentioned above are all aimed at optimizing the statistical tool of TVLA. However, it should be noted that each optimization method only addresses certain shortcomings of TVLA. Consequently, there is currently no optimal statistical tool available that can effectively solve all shortcomings
associated with the \( t \)-test used in TVLA. Therefore, when conducting leakage detection, it is essential to select a statistical tool that is appropriate for the specific environment. If addressing environmental noise, the paired \( t \)-test is recommended in place of the \( t \)-test. For horizontal and multivariate leakage, the Hotelling \( T^2 \)-test or DL-LA are suggested. Furthermore, for multi-group traces or non-mean statistical moment leakage, the \( \chi^2 \)-test, ANOVA, and KS test are recommended alternatives to the \( t \)-test. Hence, it is highly recommended to select the appropriate statistical tool based on the characteristics of the detection environment and the nature of the leakage. This ensures accurate and reliable results.

**Table 1. The optimizations of TVLA’s statistical tool.**

<table>
<thead>
<tr>
<th>Tool</th>
<th>For TVLA's Shortcoming</th>
<th>The Comparison Result with ( t )-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>The paired ( t )-test [34]</td>
<td>The environmental noise negatively affects the results of TVLA.</td>
<td>The paired ( t )-test performs better than the ( t )-test in a noisy environment.</td>
</tr>
<tr>
<td>( \chi^2 )-test [35]</td>
<td>TVLA has only two classifications; the detection results rely on the mean statistical moment.</td>
<td>When the leakage does not occur on the mean statistical moment, the ( \chi^2 )-test is better than the ( t )-test.</td>
</tr>
<tr>
<td>KS test [37]</td>
<td>The detection results rely on the mean statistical moment.</td>
<td>When the leakage does not occur on the mean statistical moment or the statistical parameters are transformed, the KS test is more robust than the ( t )-test.</td>
</tr>
<tr>
<td>Hotelling ( T^2 )-test [36]</td>
<td>TVLA cannot be used for multivariate leakage; TVLA is based on an independence assumption.</td>
<td>For multivariate leakage, compared with the ( t )-test, the Hotelling ( T^2 )-test can improve the detection efficiency.</td>
</tr>
<tr>
<td>ANOVA test [23]</td>
<td>TVLA has only two groups.</td>
<td>When the traces are divided into more groups, the detection efficiency of the ANOVA test is better than the ( t )-test.</td>
</tr>
<tr>
<td>DL-LA [14]</td>
<td>TVLA is not suitable for multivariate, horizontal leakage, and unaligned power traces.</td>
<td>For the multivariate, horizontal leakage, or unaligned power traces, DL-LA is better than the ( t )-test.</td>
</tr>
</tbody>
</table>

4.2.2. The Optimization of the Leakage Assessment Process

In TVLA, the efficiency and accuracy of detection results depend on the leakage assessment process. With this in mind, the researchers have proposed a series of suggestions to accelerate the detection process [32,33] or proposed a novel leakage assessment process [34]. This section primarily scrutinizes the optimization methods of the assessment process.

1) The optimization of TVLA’s detection process

Melissa Azouaoui studied the literature on leakage assessment and considered the leakage assessment process of TVLA as the combination or iteration of three steps: measurement and preprocessing, leakage detection and mapping, and leakage exploitation, and tried to find whether there are optimal guarantees for these steps in [33].

For measurement and preprocessing, the setting up of measurement devices depends on the expertise [65]. The preprocessing is also similar, and currently the main methods include filtering the noise [64,66] and aligning the power traces [67,68]. The best methods of setting up the measurement devices and preprocessing should be as open and repeatable as possible. Although there are some methods for setting up the measurement devices in FIPS 140 and ISO, there is currently no guaranteed optimal approach for measurement and preprocessing.

For leakage detection and mapping, the statistical hypothesis is commonly employed for comparing the distribution or statistical moment. Despite the existence of numerous methods for leakage detection, consensus on their fairness and optimality has yet to be reached. Moreover, the “budget” of traces plays a significant role in leakage detection, yet there is presently no established threshold for the optimal number of “budget” traces [29].
For leakage exploitation, it is typically divided into three stages: modeling, information extraction, and information processing. However, during the modeling stage, the evaluator utilizes the traces to estimate the optimal model based on the implementation. Nevertheless, as the number of shares increases in the mask scheme, the cost increases and the independence of samples affects the modeling phase. Currently, obtaining the optimal model remains an unresolved issue, and there is no optimal method.

During the actual leakage assessment, there are risks associated with all the aforementioned steps, and currently, there is no guarantee for an optimal leakage assessment process.

(2) A novel framework for explainable leakage assessment

Due to the unavailability of leakage detection results in TVLA, the current approach consists of utilizing a specific attack to verify the detection outcomes. Based on this, Gao Si and Elisabeth Oswald introduced a novel leakage assessment process in [34], referred to as the “the process of Gao Si” in this paper. The leakage assessment process of [34] is outlined below.

Step 1: Non-specific detection via key-dependent models.

Consider the two nested key-dependent models: The full model \( L_{cf} \) fits a model as a function of the key \( K \) to the observed data. The null model \( L_0 \) now contains only a constant term, which represents the case where there is no dependency on \( K \).

\[
L_{cf}(K_c) = \sum_j \beta_j \mu_j(K_c), j \in \left[0, 2^{16}\right],
\]

\[
L_0(K) = \beta_0,
\]

where the coefficients \( \beta_j \) are estimated from the traces via least square estimation.

The F-test is used to test \( H_0 \) (both \( L_{cf} \) and \( L_0 \) models explain the observed data equally well) versus \( H_1(L_{cf} \text{ explains the data better than } L_0) \). If the F-test finds enough evidence to reject \( H_0 \), we conclude this point’s leakage relies on \( K_c \), because \( K_c \) is a part of \( K \). Consequently, the measurement depend on \( K \).

Step 2: Degree analysis

Further restricting the degree of the key-dependent model, we determine how large the key guess is required to be to exploit an identified leakage. We obtain the model

\[
L_{cr}(K_c) = \sum_j \beta_j \mu_j(K_c), j \in \left[0, 2^{16}\right], \text{deg}(\mu_j(K_c) \leq g)
\]

The F-test is used again to test \( H_0 \) (\( L_{cr} \) and \( L_{cf} \) explain the data equally well) versus \( H_1 \) (\( L_{cf} \) explains the data better than \( L_{cr} \)).

The goal of the F-test is to have \( L_f(X) = \sum_j \beta_j \mu_j(X), j \in T_f \) with \( L_r(X) = \sum_j \beta_j \mu_j(X), j \in J_f \in T_f \). The statistical moment of the F-test is as follows.

\[
|F| = \frac{\text{RSS}_0 - \text{RSS}_f}{\frac{n_f - n_r}{\text{RSS}_f / N - n_f}}.
\]

where \( \text{RSS} = \sum_{i=1}^N \left( y^{(i)} - \tilde{L}(x^{(i)}) \right)^2 \), \( y^{(i)} \) represents the measurements of \( x^{(i)} \), and \( \tilde{L}(x^{(i)}) \) is the \( L_{cf} \) or \( L_r \) of \( x^{(i)} \). \( n_f \) and \( n_r \) are the sample size of \( L_f \) and \( L_r \). \( N \) is the sample size of \( L \). The statistical moment \(|F|\) of the F-test obeys the \( F \)-distribution with the freedom degree of \((n_f - n_r, N - n_f)\). The threshold of the \( F \)-test is \( F_{th} = Q_F(d_{f1}, d_{f2}, 1 - \alpha) \), where \( d_{f1} = n_f - 1 \), \( d_{f2} = N - n_f \), and \( Q_F \) is the quantile function of the central \( F \)-distribution.

If \( F \geq F_{th} \), the null hypothesis \( H_0 \) is rejected, \( L_f(X) \) explains the data better than \( L_r(X) \), and the leakage contains the information of the key. If there is enough evidence to reject \( H_0 \), we obtain that a model with only \( g \) or fewer key bytes suffices to explain the
measurements. By successively reducing \( g \), we can therefore determine the maximum key guess that is required to explain the side channel measurements.

**Step 3: Subkey identification**

By using the technique of further restricting the reduced model, we can narrow down precisely which specific key bytes are required to explain the identified leakage.

**Step 4: Converting to specific attacks**

If an evaluation regime requires an evaluator to demonstrate an actual attack targeting an identified leakage point, a relatively straightforward connection to a concrete profiled attack can be established.

In summary, although optimization methods exist for the measurement process and preprocessing process, there is currently no optimal leakage detection process. In actual leakage detection, the detection process of TVLA is still used to detect leakages. The process of Gao Si is a new detection process to demonstrate that the discovered leakages are key-related and can be exploited by attacks. The approach is a small step towards establishing precise attack vectors for confirmatory attacks.

### 4.2.3. The Optimization of TVLA’s Decision Strategy

(1) **The decision strategy of HC**

For long traces, the detection result is obtained through using the \( \min\,P \) strategy in TVLA. This strategy relies solely on the minimum \(|p|\) value to make a decision about leakage, disregarding all other \( p \)-values. Ding, A.A proposed the higher critical (HC) strategy in [35], which takes into account the information from all \( p \)-values.

The null hypothesis \( H_0 \): there is no leakage point in the trace. The alternative hypothesis \( H_1 \): there is at least one leakage point in the trace. Let the length of traces be \( n_l \), there are \( n_l \) \( p \)-values, denoted as \( p(1), \ldots, p(n_l) \), and the HC strategy is as follows.

**Step 1:** Sorting \( p \)-values in ascending order: \( p(1) \leq p(2) \leq \ldots \leq p(n_l) \).

**Step 2:** Calculating the normalized distance \( \hat{HC}_{n_l,i} \) of the \( p \)-value, and the normalized distance of the HC strategy is the Formula (33).

\[
\hat{HC}_{n_l,i} = \frac{\sqrt{n_l(i/n_l - p(i))}}{\sqrt{p(i)(1-p(i))}}
\]  

(33)

**Step 3:** Calculating the statistical moment of the HC strategy with (34).

\[
\hat{HC}_{n_l,max} = \max_{1 \leq i \leq n_l} \hat{HC}_{n_l,i}
\]

(34)

**Step 4:** Comparing the statistical moment \( \hat{HC}_{n_l,max} \) with the threshold \( th_{HC}^{n_l,a} \) of significance level \( a \). If \( \hat{HC}_{n_l,max} > th_{HC}^{n_l,a} \), then the null hypothesis is rejected, indicating the presence of side-channel leakages. The threshold \( th_{HC}^{n_l,a} \) represents the \( 1 - a \) quantile of the statistical moment \( \hat{HC}_{n_l,max} \) under the null hypothesis. For large \( n_l \), the threshold \( th_{HC}^{n_l,a} \) can be approximated using the connection to a Brownian bridge, such as the calculation formula provided in Li and Siegmund [68].

In summary, the HC strategy can combine multiple leakage points to enhance the efficiency of leakage detection. Thus, it can serve as a viable alternative to TVLA’s \( \min\,P \) strategy.

### 4.3. The Summary of TVLA’s Optimization Schemes

Researchers have proposed various optimization schemes for TVLA, which primarily aim to address its inherent limitations. Currently, there is no comprehensive and universally applicable statistical tool and detection process that can effectively address all the identified limitations. Suitable detection methods for different detection purposes and conditions are summarized in Figure 4.
Therefo
Therefore, assessors should perform the following process in actual leakage detection. Firstly, select the leakage detection process based on the purpose of detection. If the aim is to only discover a side-channel leakage, the recommended process is TVLA. If the aim is to detect and utilize the leakages, it is recommended to choose the process of Gao Si.

Secondly, if the TVLA process is chosen, it is recommended to select an appropriate statistical tool based on the evaluator’s prior knowledge of the device. If the evaluator has no prior knowledge about the device, a $t$-test is used initially to detect whether there is a univariate, first-order leakage. If a leakage is detected, the leakage detection process is stopped. Otherwise, the $\chi^2$-test, KS test and ANOVA test are used to detect univariate, high-order leakages, while the Hotelling $T^2$-test and DA-LA are used to test for the presence of multivariate or horizontal leakage. If the evaluator has prior knowledge of the device, the leakage type (univariate or multivariate) is determined based on this knowledge. Then, the detection environment (high noise or low noise) and alignment of power traces are determined. For univariate, first-order leakage (mean statistical moment) and low-noise environments, the $t$-test is more efficient. For univariate, first-order leakage and high-noise environments, the paired $t$-tests have better detection efficiency. For univariates where the leakage does not occur at the mean statistical moment, the $\chi^2$-test, KS test, and ANOVA test have better detection performance than the $t$-test. For multivariate and horizontal leakages, it is recommended to choose the Hotelling $T^2$-test and DA-LA, with DA-LA being more effective when the traces are unaligned.

Finally, in the determination stage, the HC strategy is recommended.

Although many leakage detection methods have been proposed, TVLA remains the mainstream detection method. Therefore, the $t$-test is generally regarded as the mainstream detection tool, with other detection tools considered supplementary. However, the current
5. Quantification of Side Channel Vulnerability

The inability of TVLA’s detection results to quantify the vulnerability of side channels raises the question of how to establish a relationship between attacking-style assessment and leakage detection-style assessment. Debapriya made an attempt to derive this relationship between TVLA and SCA. Additionally, SR can be calculated directly from TVLA, effectively connecting CC and FIPS 140-3 based on the provided intermediate variables and leakage model in [22].

The derivation process of the relationship between TVLA and SR is as follows. Let \( L = f(X, k) \) be the normalized leakage model, where \( E(L) = 0, \text{Var}(L) = E(L^2) = 1 \). \( Y \) represents the measurements, and it can be defined as \( Y = cL + \mathcal{N} \), where \( c \) is the scale factor and \( \mathcal{N} \sim \mathcal{N}(0, \sigma^2) \) represents the noise. Taking S-box as an example, the \( n \)-bit Hamming weight model can be expressed as \( f(X, k) = \frac{2}{n} (\text{HW}(sbox(X \otimes k)) - \frac{n}{2}) \) (*).

Firstly, link TVLA and NICV. We obtained \( \text{NICV}_2 = \frac{n}{(TVLA)^2 + \frac{n}{2}} \). If \( n_2 = n_1 = \frac{n}{2} \), \( \text{NICV}_2 = \frac{1}{(TVLA)^2 + 1} \). If the side-channel traces are divided into \( q \) groups, then

\[
\text{NICV}_q = q - \frac{1}{q} \sum_{i=1}^{q} \text{NICV}_i^q
\]

where \( \text{NICV}_2 = \frac{\sum_{i=1}^{q} n_i (\mu_i - \mu)^2}{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu)^2} \), \( n_i = n - n_j \).

Secondly, link SNR and NICV. The form of SNR, NICV, and TVLA under the leakage model (*) are \( SNR = \frac{\text{Var}(E(Y|X))}{E(\text{Var}(Y|X))} \), \( SR = \frac{c^2}{\sigma^2} \), \( NICV = \frac{\text{Var}(E(Y|X))}{\text{Var}(Y)} \), and \( SR = \frac{1}{1 + \frac{c^2}{\sigma^2}} \). Then

\[
\text{NICV} = \frac{1}{SNR + 1} = \frac{e^{l(X,q)}}{\sqrt{e^{2} + 2\sigma^2}}
\]

Finally, export SR using SNR. The relationship between SNR and SR is

\[
SR = \Phi_{\frac{1}{K+1}SNR(K^{*}-\|\|T)} \left( \sqrt{\frac{1}{2}} \sqrt{\text{SNR}} \right),
\]

where \( \| = (\|c, k_{g_1}\|, \ldots, \|c, k_{g_{2^n-1}}\|)^T \), \( \|c, k_c\| = E \left( (l(X, k_c) - l(X, k_{g_1}))^2 \right) \).

\[
K = \begin{pmatrix}
\|c, k_{g_1}\| & \ldots & \|c, k_{g_1}, k_{g_{2^n-1}}\| \\
\vdots & \ddots & \vdots \\
\|c, k_{g_{2^n-1}}, k_{g_1}\| & \ldots & \|c, k_{g_{2^n-1}}, k_{g_{2^n-1}}\|
\end{pmatrix},
\]

\[
\|c, k_{g_1}, k_{g_1}\| = E \left( (l(X, k_c) - l(X, k_{g_1})) (l(X, k_c) - l(X, k_{g_1})) \right),
\]

\[
K^{**} = \begin{pmatrix}
\|^{**}(c, k_{g_1}, k_{g_1}) & \ldots & \|^{**}(c, k_{g_1}, k_{g_{2^n-1}}) \\
\vdots & \ddots & \vdots \\
\|^{**}(c, k_{g_{2^n-1}}, k_{g_1}) & \ldots & \|^{**}(c, k_{g_{2^n-1}}, k_{g_{2^n-1}})
\end{pmatrix}
\]

\[
\|^{**}(c, k_{g_1}, k_{g_1}) = 4E((l(X, k_c) - E(l(X, k_c)))^2) (l(X, k_c) - l(X, k_{g_1}))(l(X, k_c) - l(X, k_{g_1})),
\]

where \( \Phi_{\|g_{1}\|}(\mu) \) is the cumulative distributive function of the multivariate normal distribution with mean vector \( \mu \) and covariance \( C \), \( k_c \) is the correct key, and \( k_{g_i} \) with \( 1 \leq i \leq 2n - 1 \).
Figure 5 presents the process of hybrid side-channel testing. The process consists of non-specific TVLA $\rightarrow$ specific TVLA $\rightarrow$ SR $\rightarrow$ evaluation results. First, the evaluator performs a non-specific TVLA on the target device, followed by the calculation of SR using specific TVLA. The side-channel vulnerabilities of the target device are assessed without the need for actual attacks. If SR is below the security limit, the device is considered safe; otherwise, it is considered vulnerable to SCA.

![Process Diagram]

Figure 5. The process of hybrid side-channel testing.

Although this method aims to establish the relationship between TVLA and SR, it provides limited bridging between CC and FIPS. However, this method is based on the assumptions regarding intermediate variables and leakage models, resulting in difficulties in practical detection for evaluators without professional knowledge of these models and values.

6. Discussion

Based on the research on leakage assessment works, we discuss leakage assessment from two perspectives: the current status and future development trend.

Regarding the current status of leakage assessment, firstly, although many leakage assessment methods have been proposed to address the specific shortcomings of TVLA (multivariate, high-order leakage, noise issues, or independence of traces), there is currently no unified standardized method to guide all assessors in conducting the assessment step by step. Secondly, in actual leakage detection, the evaluator or agencies often spend a significant amount of time collecting the traces, regardless of the detection method used. The goal is to evaluate the security of products and identify potential vulnerabilities and provide further design to achieve the required security level. However, current leakage detection focuses solely on discovering leakages and does not uncover vulnerabilities or
guide the design process. From an investment cost perspective, the income obtained from leakage detection is low.

In terms of the future development trend of leakage assessment, researchers are currently attempting to link the non-specific leakage detection results with key guessing in order to address the issue of unusable detection results. This approach would allow the detection results to reveal information about the key or causes of leakage, which can aid designers in constructing attacks or defenses. Therefore, we anticipate the formation of a unified and fair method for security assessment. Additionally, in recent years, deep learning technology has been applied to leakage detection. Compared to the traditional leakage assessment technology, the deep learning leakage assessment offers simplicity and high level of statistical confidence. However, it lacks the ability to provide superior security metrics. Therefore, the further exploration is needed on how to effectively apply the deep learning technology to side-channel leakage assessment.

7. Conclusions

In this paper, we conducted a comprehensive study on the leakage detection-style assessment. These leakage detection-style assessment methodologies can be classified into two categories: TVLA and its optimizations. We identified the drawbacks of TVLA and categorized the optimization schemes aimed at addressing these limitations into three groups: statistical tool optimization, detection process optimization, and decision strategy optimization. We gave succinct descriptions of the leakage detection assessment motivations and detection processes and compared the efficiency of different optimization schemes. Based on our classification and summary, we concluded that there is no single optimal scheme that can effectively address all the shortcomings of TVLA. Different optimization schemes are proposed for specific purposes and detection conditions. We summarized the purposes and conditions for all TVLA optimizations and proposed a selection strategy for leakage detection-style assessment schemes. According to the selection strategy, the leakage detection process should be chosen based on the specific detection purpose. For discovering side-channel leakages, TVLA is recommended, while the detection process of Gao Si is suitable for discovering and utilizing leakages. Additionally, the appropriate statistical tool should be selected. T-tests and paired t-tests are for detecting univariate, first-order leakage. In low-noise environments, the t-test is more suitable, whereas in high-noise environments, the paired t-test demonstrates better detection efficiency. If the leakage does not occur at the mean statistical moment, the χ²-test, KS test and ANOVA test outperform the t-test for detecting univariate, high-order leakages. For testing multivariate or horizontal leakages, the Hotelling T²-test and DA-LA are recommended. If the traces are unaligned, DA-LA is more effective. Lastly, the HC strategy is recommended for the determination stage. Based on the current status of leakage detection-style assessment, we discuss the development trend of leakage detection. Researchers are increasingly interested in linking leakage detection with key guessing to address the issue of unusable detection results. The aim is to establish a unified and fair method for security assessment.

Author Contributions: Y.W. and M.T. made substantial contributions to the conceptualization and methodology of the review. Y.W. also contributed to the writing, review, and editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key R&D Program of China No. 2022YFB3103800 and the National Natural Science Foundation of China No. 61972295.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
References


2. Cao, S.; Fan, L. NSA's top backdoor has been exposed by Chinese researchers. *Glob. Times* **2022**. [CrossRef]


20. FIPS 140-3; Security Requirements for Cryptographic Modules. NIST: Gaithersburg, MD, USA, 2019.


