

Detecting Misuses of Security APIs: A Systematic Review

ZAHRA MOUSAVI, CREST - The Centre for Research on Engineering Software Technologies, University of Adelaide, Cyber Security Cooperative Research Centre, CSIRO/Data61, Australia

CHADNI ISLAM, Queensland University of Technology, Australia

M. ALI BABAR, CREST - The Centre for Research on Engineering Software Technologies, University of Adelaide, Australia

ALSHARIF ABUADBBA and KRISTEN MOORE, CSIRO/Data61, Australia

Security Application Programming Interfaces (APIs) play a vital role in ensuring software security. However, misuse of security APIs may introduce vulnerabilities that can be exploited by hackers. API design complexities, inadequate documentation and insufficient security training are some of the reasons for misusing security APIs. In order to help developers and organizations, software security community have devised and evaluated several approaches to detecting misuses of security APIs. We rigorously analyzed and synthesized the literature on security APIs misuses for building a body of knowledge on the topic. Our review has identified and discussed the security APIs studied from misuse perspective, the types of reported misuses and the approaches developed to detect misuses and how the proposed approaches have been evaluated. Our review has also highlighted the open research issues for advancing the state-of-the-art of detecting misuse of security APIs.

CCS Concepts: • **Security and privacy** → **Software and application security**.

Additional Key Words and Phrases: Security API, Secure Software Development, API Misuse, Misuse Detection

ACM Reference Format:

Zahra Mousavi, Chadni Islam, M. Ali Babar, Alsharif Abuadbba, and Kristen Moore. 2023. Detecting Misuses of Security APIs: A Systematic Review. In *Proceedings of ACM Computing Survey (CSUR)*. ACM, New York, NY, USA, 35 pages.

1 INTRODUCTION

There has been dramatic increase in successful cybersecurity attacks that usually result in high profile data breaches. According to the latest data breach report by IBM and the Ponemon Institute, the average cost of a data breach in 2022 was US\$ 4.35 million indicating a 12.7% rise compared to the average cost of \$3.86 million documented in the 2020 report [1]. A majority of successful data breaches can be attributed to software vulnerabilities [2], for example Equifax data breach in USA [3] or Optus data breach in Australia [4]. That is why there is an increased emphasis on carefully considering and addressing security concerns during the software development life cycle. Developers often rely on security Application Programming Interfaces (APIs) developed by security experts in order to embed security functionalities (e.g., authentication, authorization, and data integrity) into their code without the need of comprehending the underlying concepts and technical details of the functionality [5]. An example of security APIs is cryptography APIs, which are widely used to ensure the confidentiality of sensitive data and make secure communications [6].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

Manuscript submitted to ACM

Whilst the availability of security APIs greatly benefits in ensuring software security, their incorrect use can have significant negative effects on the security of a software application [7][8][9]. Several studies have shown that misuse of security APIs is widespread. Based on a study of ten thousand Android applications, Krüger et al. [6] showed that approximately 95% of the applications suffered from at least one cryptography API misuse. Similarly, a study of over two thousand open-source Java projects on GitHub showed that 72% of the projects have at least one cryptography API misuse [10]. Examples of correctly using a security API of cryptography can be the selection of algorithms that are deemed secure, generating a key using a secure random number generator with a secure seed and avoiding the use of a hardcoded constant key for encryption that attackers may be able to access [11].

Misuses of security APIs can be attributed to several reasons such as lack of developers' security knowledge and skills, lack of usability consideration in APIs design or lack of sufficient documentation of security APIs [12][13][14]. Given most of these reasons for misuse of security APIs are related to the users, i.e., software developers, it is understandable that developers often struggle to properly integrate security APIs into their code without having highly developed knowledge and skills in software security. Several studies report that generally developers lack appropriate training and skills in secure software development of which the use of security APIs is an integral part [14][15]. Lack of sufficient training and skills in security aware software development practices and unavailability of easily accessible and understandable documentation of security APIs may encourage developers to rely on readily accessible information sources on open source repositories and forums such as SO [14][16]. However, the answers provided on social media forum may lead to actions that can introduce vulnerabilities in software [17].

Irrespective of the cause of misuse of security APIs, it is important to effectively and efficiently detect and correct them for ensuring the security of software applications using security APIs [18]. Given the increasing realization of the potentially devastating consequences of misuse of security APIs, there has been significant interest in devising and evaluating effective and efficient approaches to detecting security API misuses. Detecting the increasing research interests in approaches to detecting security APIs misuses, the relevant literature on this topic is dispersed without systemic analysis and synthesis. We assert that a systematic survey of the available peer-reviewed literature can assist researchers and practitioners to better comprehend the reported types of misuses of security APIs, the proposed approaches to detect misuses and their limitations, and the research gaps in the existing needs and solutions to detecting and correcting misuses of security APIs. To fill this research gap, we conducted a Systematic Literature Review (SLR) of the existing studies, focused on detecting misuses of security APIs, with the aim of providing a coherent body of knowledge on this topic. The main **contributions** of this research study are:

- A large scale survey of the literature on misuse detection for security APIs using a systematic review method.
- An in-depth discussion on the security APIs studied from misuse perspective.
- A taxonomic analysis of the existing approaches to detecting misuses of security APIs.
- A critical rundown of the strategies, metrics, and benchmarks used for evaluating the proposed approaches.
- A set of open issues that can form the future research agenda for the devising and evaluating approaches to detecting misuses of security APIs.

The remainder of this paper is structured as follows. Section 2 provides an overview of security APIs and misuses. Section 3 describes the scientific methodology used to conduct this SLR. Our findings in terms of security APIs and misuses are presented in Sections 4 and 5, respectively. Section 6 analyzes the techniques used for detecting misuses, while Section 7 presents the evaluation methods and their results. Section 8 discusses open issues and future research directions. Section 9 discusses the threats to the validity of our findings. Finally, Section 10 concludes the paper.

2 PRELIMINARIES

This section presents an overview of security APIs and potential misuses that developers may make while using them.

2.1 Security API

An API is essentially a set of programming instructions that allows software components to interact with each other, making it easier for developers to build complex software systems [19]. APIs can provide various functionalities such as data management, payment processing, or messaging, that can be integrated with software systems. Security APIs are a subset of APIs that provide developers with security functionalities, such as confidentiality, data integrity, authentication, and authorization [20]. Confidentiality is the process of protecting sensitive information from unauthorized disclosure [21]. Data encryption, provided by cryptography APIs, is a primary means to ensure the confidentiality of sensitive data. Similarly, watermarking and steganography APIs can be used for information hiding and confidentiality. Data integrity mechanisms ensure the accuracy, trustworthiness, and validity of data by protecting it from unauthorized changes throughout its life cycle. Cryptography and SSL/TLS APIs are among the popular security APIs that provide security functionalities to ensure data integrity [6].

Authentication is the process of verifying the identity of legitimate users or systems before granting data access [21]. Authentication APIs enable developers to integrate this functionality into their applications, allowing them to verify the authenticity of data access requests. Authorization is the process of access control, which specifies access rights and privileges to resources [21]. OAuth APIs [22] are used for both authentication and authorization purposes, enabling users to grant access to their resources and data to third-party applications without exposing their credentials.

2.2 Misuse of Security APIs

APIs only operate properly when certain constraints on their inputs, outputs, and invocation context are met which is known as API specification. Violation of API specifications is called misuse. In the case of security, misuse of APIs may lead to severe security consequences. Therefore, it is essential for developers to strictly adhere to the most up-to-date specifications for using security APIs correctly. For example, to establish secure communication using SSL/TLS API, a client must verify a server hostname. Figure 1 illustrates a misuse of the specification of that API. It shows how the misuse of SSL/TLS API by allowing all hostnames results in the leakage of user credentials via a Man-in-the-Middle attack. Attackers can impersonate a valid server and intercept the communication between a user and applications, gaining unauthorized access to the user’s personal information.

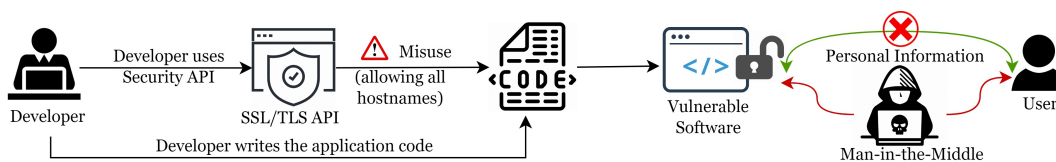


Fig. 1. A misuse of SSL/TLS API leading to the leakage of user credentials

The complexity of API designs poses a major hurdle when it comes to using security APIs correctly, often making it difficult for developers to comprehend and implement them effectively [12]. Using security APIs accurately can be further complicated by inadequate or poor documentation that fails to provide explicit examples of proper usage or even includes insecure code samples. For instance, some insecure samples were found among the code examples provided by some OAuth API providers for helping developers with implementing OAuth connections [23]. Poor API documentation often causes developers to rely on forum posts to learn how to use a particular API, which leads to

copy-pasting incorrect suggestions that contain misuse instances [17]. On the other hand, poor default configurations for some APIs further exacerbate the situation. For example, *Java Cryptography Architecture* (a popular cryptography API) uses ECB mode by default for the “AES” algorithm which is not secure[6].

Developers usually lack cybersecurity training and may prioritize enhancing other features over security [24]. In today’s fast-paced software development environment, there is often pressure to release software quickly, leading developers to take shortcuts and not fully test their code for security vulnerabilities, including security API misuses. Furthermore, the threat landscape for security APIs is continuously evolving, with new attack techniques and vulnerabilities being discovered regularly. As a result, developers may find it challenging to keep up with the latest updates and best practices for using security APIs. All these challenges contribute to the widespread misuse of security APIs, emphasizing the necessity to investigate state-of-the-art techniques for detecting misuse in this domain.

3 RESEARCH METHODOLOGY

We conducted an SLR to gain insight into misuse detection approaches for security APIs. SLR is broadly adopted as a research methodology in Evidence-Based Software Engineering [25] as it provides a reliable, rigorous, and auditable technique for assessing and interpreting a research topic [26]. We followed the SLR guideline provided by Kitchenham et al. [26] and formulated four Research Questions (RQs) with corresponding motivations, as detailed in Table 1 to guide our analysis. The steps of our review protocol are elaborated in Subsections 3.1- 3.4.

Table 1. Research questions addressed in this study

Research Questions	Motivation
RQ1: Which security APIs have been studied in the context of misuse detection?	To identify security APIs studied by researchers for misuse detection and shed light on their concept and functionality. It allows us to understand which security APIs have been considered more critical in this area or which are disregarded.
RQ2: Which security API misuses have been studied in the literature?	To provide insight to practitioners and researchers on security API misuses targeted by existing misuse detection literature and the security threats they can cause.
RQ3: What types of techniques have been used to detect security API misuses?	To investigate state-of-the-art approaches that are used to detect misuses of security APIs and give detailed information on their design and implementation.
RQ4: What are strategies used to evaluate the performance of misuse detection techniques?	To investigate procedures in the literature for measuring the performance of approaches, including employed datasets, benchmarks, and metrics.

3.1 Search Strategy

We followed the guideline provided by Kitchenham et al. [26] to develop our search strategy, ensuring that we obtain the highest number of relevant studies. The search strategy is comprised of the following steps.

3.1.1 Search Method. We applied an automated database search method [26] to digital search engines and databases to obtain relevant studies. We used Scopus Digital Library (DL) as the primary source, which is the largest academic literature database, indexing over 5,000 publishers worldwide, including relevant sources like Elsevier and Springer [27, 28]. To complement Scopus results, we also used the two prominent academic DLs – IEEE Xplore and ACM DL [29].

3.1.2 Search String. We crafted a comprehensive search string following the guidelines presented by Kitchenham et al. [26]. We initiated our search using four main keywords: “security”, “API”, “misuse”, and “detection”. To broaden our search, we also considered synonyms for these terms. We reviewed titles, abstracts, and keywords from some relevant papers to ensure we captured associated synonyms. Synonyms that returned an excess of irrelevant results, such as “flaw”, were excluded from our search string. We conducted a series of pilot searches to ensure the inclusion of relevant papers that we were already aware of. Ultimately, we organized the keywords and their pertinent synonyms into four categories, which are shown in Table 2. We used the union (AND) of the categories to conduct searches in titles, abstracts, and keywords of papers on Scopus, IEEE Xplorer, and ACM DL.

Table 2. Categories of key terms used for defining search string

Category	Synonyms and Relevant Terms
Security	<i>secur* OR crypto*</i>
API	<i>api OR librar* OR interface</i>
Misuse	<i>misuse OR incorrect OR insecure OR vulnerabilit*</i>
Detection	<i>detect* OR "static analysis" OR "dynamic analysis" OR "program analysis" OR "code analysis"</i>

3.2 Inclusion-Exclusion Criteria

We assessed the papers retrieved from databases by applying specific criteria for the inclusion of relevant studies and the exclusion of out-of-scope papers. The inclusion and exclusion criteria were defined and aligned with our research objectives and RQs. Moreover, we adopted a venue assessment criterion, similar to methods used in previous studies [28] [30], to ensure that we only included high-quality papers. We removed papers published in low-quality venues, including unranked conferences and workshops according to the CORE ranking¹, and unranked journal venues according to the Scimago database². Both databases employ meticulous and comprehensive evaluation methodologies considering various factors to assess the quality and impact of venues. By utilizing these databases, we were able to identify high-quality papers effectively. Table 3 presents the final list of inclusion and exclusion criteria for this SLR.

Table 3. Inclusion and Exclusion Criteria

Inclusion Criteria
I1: Papers that address misuse detection for security APIs, including papers that use either an automated or manual approach or rely on existing tools to verify the usage of security APIs.
I2: Papers published in peer-reviewed venues that are ranked by CORE or Scimago.
I3: Papers written in English and their full text are accessible.
Exclusion Criteria
E1: Papers that target detecting vulnerabilities in the internal design or implementation of security APIs, not misuses.
E2: Papers that target detecting misuses of generic APIs.
E3: Short papers less than six pages.
E4: Book chapters, dissertations, and non-peer-reviewed publications (e.g., keynotes, editorials, tutorials, and panel discussions).

3.3 Selection of the Primary Studies

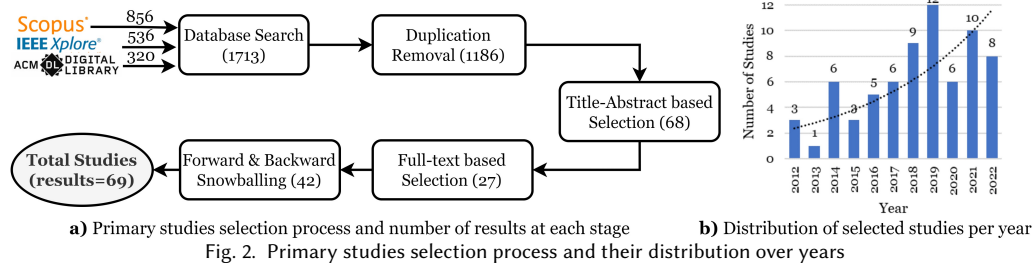
Figure 2.a illustrates the different phases involved in the study selection process and the number of studies obtained at each stage. In November 2022, we executed the search string in our data sources without applying any time limit to the publication year, resulting in 1,713 studies. Subsequently, we applied the following steps to refine our search: duplication removal, selection based on title and abstract, and selection based on full text according to our inclusion-exclusion criteria outlined in Table 3. Furthermore, we used forward and backward snowballing techniques [31] to ensure the maximum number of relevant papers were included in our review. This involves examining the citations and references of the selected papers to identify any missing relevant papers. Finally, 69 papers were included in our SLR, and their details can be found in our online appendix [32]. Each paper in the review is assigned a unique identifier (S#)

3.4 Data Extraction and Synthesis

We developed a Data Extraction Form (DEF) that comprises 14 data items essential for addressing our RQs, which we elaborated in our online appendix [32]. Data items (D1-D6) include demographic information such as title, author, venue, publication year, and publisher. To simplify the analysis of the extracted data relevant to our RQs, we categorized the data items into the following groups: RQ1 (D7-D8: security APIs, language), RQ2 (D9-D10: misuses, consequences), RQ3

¹<http://portal.core.edu.au/conf-ranks/>

²<https://www.scimagojr.com/journalrank.php>



(D11-D14: technique, modeling input data type, testing input data type, output type), and RQ4 (D15-D18: evaluation strategy, evaluation metrics, dataset, misuses reported). A pilot study was conducted on 12 papers to refine the DEF for capturing the necessary information in the most effective and summarized form.

Data Synthesis: We used descriptive statistics to analyze demographic information data items, while thematic analysis was used to analyze RQ-relevant data items. To conduct the thematic analysis, we followed the steps outlined in the guideline of study [33]. Firstly, we familiarized ourselves with the data by reading and examining the extracted data. Next, we generated initial codes to capture security APIs, misuses, detection techniques, and evaluation methods. Then, we searched for themes and generated potential themes for each data item by merging the corresponding initial codes based on their similarities. We reviewed the themes and mapped them iteratively to ensure all codes and themes were accurately allocated. Finally, we reviewed the synthesized results for each RQ and resolved any disagreements through regular meetings to finalize the answers to RQs.

3.5 Studies Distribution

Figure 2.b illustrates the distribution of 69 primary studies over the years. It is noteworthy that no relevant studies were found prior to 2012, and since then, there has been an upward trend in publications toward 2022. The rising number of papers highlights a growing interest from the research community in the field of misuse detection for security APIs. This upward trend indicates the expanding threat landscape of misuses and emphasizes the crucial demand for effective approaches to detect and mitigate them.

4 RQ1: SECURITY APIS

This section presents our findings regarding RQ1, which focuses on the security APIs researchers have studied for the purpose of misuse detection. Our analysis revealed that the state-of-the-art studies are focused on investigating and analyzing the usage patterns of six security APIs, as listed in Table 4. Among these, the **cryptographic primitives** API has been the most extensively studied API, with 43 studies evaluating its usage. Researchers have also shown interest in investigating the **SSL/TLS** API, with 26 studies explicitly focusing on this API. For implementing authentication and authorization functionality, 9 studies evaluated the usage of **OAuth** APIs, while one study focused on the **Spring** framework, and another study is dedicated to the **Fingerprint** API. Additionally, one study has investigated how **SafetyNet Attestation** API has been used in Android applications to implement app/device integrity.

Further investigation into the programming languages of these APIs, it was discovered that the majority of the research focused on security APIs in Java. This is possibly due to the complexity of the design of Java APIs [S20, S60], highlighting the demand for the development of effective approaches to detect misuses in Java APIs. Moreover, the popularity of Java and its extensive usage in Android app development have further emphasized the need to mitigate the misuse of Java security APIs. This issue is particularly salient given the widespread use of mobile devices for accessing

Table 4. Security APIs and their mappings with primary studies (number of primary studies indicated in parentheses)

APIs	Functionality	Language	Instances	Study Refs
Cryptographic Primitives (43)	Confidentiality Data Integrity Authentication	Java	Java Cryptography Architecture (JCA), Java Cryptography Extension (JCE), BouncyCastle, Jasypt, Keyczar, GNU Crypto, SunJCE, BouncyCastle, SpoungyCastle, LP11	S1, S2, S3, S4, S5, S7, S10, S11, S12, S15, S16, S17, S18, S20, S23, S27, S29, S31, S32, S34, S35, S36, S40, S41, S42, S43, S44, S45, S46, S55, S58, S61, S63, S64, S65, S66
		Python	PyCrypto, PyNaCl, M2Crypto, cryptography.io, Keyczar, ucryptolib	S20, S59, S60
		C/C++	CommonCrypto, Libsodium, Nettle, TomCrypt, LibTomCrypt, Libgcrypt, WolfCrypt	S6, S13, S28
		JavaScript	WebCrypto APIs	S56
		Go	Go cryptographic APIs	S62
SSL/TLS (26)	Confidentiality Data Integrity Authentication	Java	Java Secured-Socket Extension (JSSE)	S1, S2, S3, S8, S14, S19, S24, S26, S27, S31, S33, S34, S39, S43, S44, S45, S46, S57, S58, S66, S67, S68
		C/C++	OpenSSL, GnuTLS, Libcrypto, Libcrypt, Cryptlib, WolfSSL	S9, S21, S22, S25, S26, S68
OAuth (9)	Authentication Authorization	-	OAuth APIs provided by service providers such as Google or Facebook	S30, S47, S48, S49, S50, S51, S52, S53, S54
Fingerprint (1)	Authentication Authorization	Java	Google Fingerprint API	S37
Spring (1)	Authentication Authorization	Java	Spring framework	S38
SafetyNet Attestation (1)	Device/App Integrity	Java	Google SafetyNet Attestation	S69

and storing sensitive information, coupled with Android being the dominant mobile operating system. Aside from Java, there were several studies focused on exploring the use of security APIs in other programming languages. Specifically, there were 8 studies dedicated to C/C++, 3 studies on Python, and one study each on JavaScript and Go (Table 4).

The following subsections provide an overview of each security API, focusing on its key components and the functionalities offered by each API.

4.1 Cryptographic primitives APIs

Cryptography is an essential component of secure software development as it plays a crucial role in maintaining confidentiality, data integrity, and authenticity. Developers often utilize APIs that implement cryptography primitives (referred to as crypto APIs, hereafter) to integrate these features into their software. Cryptography primitives are fundamental building blocks of cryptography. They consist of low-level functions including (i) *symmetric encryption*, (ii) *asymmetric encryption*, (iii) *hash and message authentication code*, (iv) *key derivation*, (v) *key storage*, and (vi) *pseudorandom number generator*.

Symmetric Encryption: Encryption algorithms, called *ciphers*, protect data confidentiality by converting *plaintext* into *ciphertext* that can only be decrypted by authorized entities. Symmetric encryption, known as private key cryptography, uses the same key for both encryption and decryption. Block ciphers are the most prevalent type of symmetric encryption that divide plaintext into fixed-size blocks and encrypt them into ciphertext blocks of the same size.

Asymmetric Encryption: Asymmetric encryption, also known as *public-key cryptography*, uses two distinct keys, a *public key* and a *private key*. The public key is used to encrypt the data, while the private key is used to decrypt it. In addition, asymmetric encryption can be used to implement *digital signatures* for ensuring authenticity in communications. To this end, the sender signs data with a private key, and the receiver verifies the signature with the sender’s public key.

Hash and Message Authentication Code: Hash functions maintain data integrity by converting input data of arbitrary length into unique and fixed-length hash values. As slight changes in the input result in completely different hashes, hash functions are effective for detecting any modification to the original data. Message Authentication Codes (MACs), while similar to hash functions, also incorporate a secret key. This key allows the sender to authenticate their identity as the the message's origin, thereby ensuring both authenticity and integrity.

Key Derivation: A Key Derivation Function (KDF) generates a cryptographic key from a *password* or *passphrase* that fulfills standards such as minimum length, entropy, and brute-force resistance. It is commonly employed in combination with *Password-Based Encryption (PBE)*. The process of key derivation through a KDF typically involves applying a hash function, using a random value, called *salt*, for an adequate number of *iterations* to prevent brute-force attacks.

Key Storage: Preserving the confidentiality and integrity of encrypted data in cryptography heavily relies on proper key storage practices. Key storage algorithms are designed to assist developers in securely storing sensitive credentials, such as key material. These algorithms require a strong *password* or *passphrase* as input to provide adequate security.

PseudoRandom Number Generator: Randomness plays a crucial role in all aspects of cryptography. Cryptography APIs offer PseudoRandom Number Generator (PRNG) functions to ensure the generated number holds the requisite level of randomness for cryptographic applications. PRNGs rely on a seed for generating random numbers that must also be random to prevent any potential predictability in the generated numbers.

Our survey covered 43 studies examining the usage of crypto APIs. Among these, 36 studies explored the usage of Java Cryptography Architecture (JCA), while Python PyCrypto API, C/C++ CommonCrypto API, JavaScript, and Go were each subject to 3, 3, 1, and 1 study respectively.

4.2 SSL/TLS APIs

Secure Sockets Layer (SSL) and Transport Layer Security (TLS) protocols are used to establish a secure channel for communication between a client and a server, protecting them from potential attacks like a Man-in-the-Middle (MitM). These protocols rely on cryptographic primitives to ensure authentication, confidentiality, and integrity of network messages. To establish and validate SSL connections, developers can utilize SSL/TLS APIs such as OpenSSL [34], which encapsulate the details and functionalities of these protocols. A critical aspect of SSL connection establishment is authenticating the server. During the SSL handshake, the server presents its *public key certificate* to the client as a means of authentication. It is essential that the client carefully verifies the authenticity of the server's certificate to ensure the security of the SSL connection. The validation process involves **certificate validation** and **hostname verification**.

In **certificate validation**, a client must carefully verify that the certificate has been issued and signed by a trusted *Certificate Authority (CA)* and has not expired or been revoked. Additionally, the client must validate the certificate chain presented by the server. This involves verifying each certificate in the chain and ensuring that each one is issued by the CA immediately above it, with a trusted root CA at the top of the chain.

In **hostname verification**, a client must verify that the *hostname* included in the certificate matches the hostname that the client is attempting to connect. This verification process prevents MitM attacks, where an attacker intercepts the communication between a client and server, and impersonates the server by sending false information to the client.

In our survey, 26 studies focused on the usage of SSL/TLS APIs, with 22 studies for Java APIs, e.g., Java Secured-Socket Extension (JSSE), and 6 studies for C/C++ APIs, e.g., OpenSSL.

4.3 OAuth APIs

Open Authorization (OAuth) [22] is a popular authorization protocol enabling end-users to grant third-party websites or applications access to their private resources stored on a remote server without sharing their credentials. This process involves three major roles: the **user** or **resource owner** who owns protected resources, the **Service Provider (SP)** that hosts the resources, and **Relying Party (RP)**, known as *client application* that uses the SP to obtain access to the user’s resources. **Access tokens** are issued by SPs to RPs with the owner’s approval for accessing protected resources.

OAuth APIs provided by SPs such as Google, Twitter, or Facebook are used by developers to authenticate users or obtain access to users’ resources through their major accounts in SPs. Although OAuth was first introduced as an authorization framework, it has been widely adopted to implement Single-Sign-On authentication, making it difficult for developers to use it properly. To address this challenge, OpenID Connect [35] was introduced as an authentication framework based on OAuth. The OAuth specification [22] defines four different protocol flows or grant types: (i) **authorization code**, (ii) **implicit**, (iii) **resource owner password credentials**, and (iv) **client credentials**.

Figure 3.a depicts the process of authorization code grant which is the the most commonly used grant type. As shown in Figure 3.a, the process begins with a user sending a request to a RP to access a remote resource (step 1). The RP then redirects the user to the SP with an *APP ID* and an optional *state* parameter to bind this request (step 2). Next, the user authenticates with the SP and grants the RP’s requested permissions (step 3). The SP issues an *authorization code* and an optional *state* parameter to the user (step 4). The user is then redirected back to the RP’s redirection endpoint, where the request is rejected if the received *state* parameter mismatches the initial one (step 5). Next, the RP sends the *authorization code* and its *secret* (established during registration with the SP) to the SP to request an access token (step 6). The SP verifies the RP app by validating the *App ID* and *app secret* and then responds with an *access token* (steps 7). With the access token the RP requests user data from the SP, which is then shared with the RP accordingly (steps 8-9).

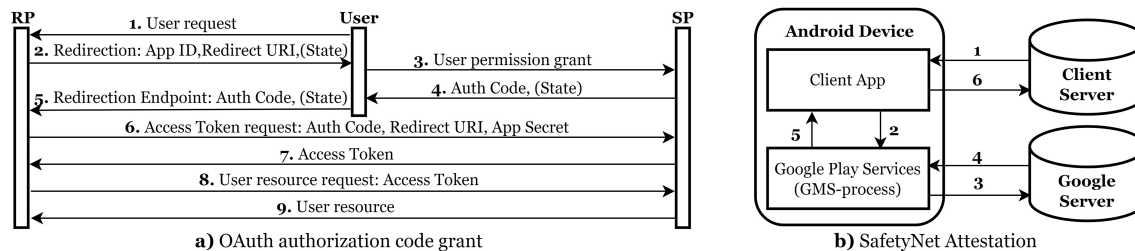


Fig. 3. (a) OAuth authorization code grant flow between user, relying party (RP) and service provider (SP) and (b) attestation process performed using Google SafetyNet Attestation API [S69]

In the implicit grant (which is simpler than the authorization code grant), in step 4, the SP directly responds with an access token instead of an authorization code, without authenticating the RP. Resource owner password credentials and client credentials grants are rarely used.

Our review included 8 studies that evaluated the usage of OAuth APIs in Android and web applications. Additionally, one study [S54] focused on both OAuth and OpenID Connect APIs for implementing authentication in Android apps.

4.4 Fingerprint APIs

Our review identified one study [S37] that investigated the usage of Fingerprint API in Android apps. With the increasing use of smartphones for security-sensitive tasks such as mobile payments and banking, ensuring a highly secure authentication process is crucial. *Two-factor authentication (2FA)* was developed to enhance security and usability

by incorporating an *additional factor* during authentication. While smartphones can serve as a second factor in 2FA, they also pose significant security threats if stolen or compromised. Modern smartphones equipped with *Trusted Execution Environments (TEE)* can securely generate and store cryptographic keys. Combining TEE with fingerprint readers for 2FA provides strong security comparable to external hardware devices such as YubiKeys [36].

Both Google [37] and OWASP [38] guidelines recommend using a fingerprint reader in conjunction with cryptographic operations for secure authentication. This involves using the fingerprint to unlock a cryptographic key protected by the TEE, rather than just recognizing the user. To interact with the fingerprint sensor and verify whether a legitimate user has touched it, four essential steps are required to follow [S37]. That begins with **generating a cryptographic key** where developers specify key properties via parameters such as setting the *user authentication required* parameter to *True* ensuring key usability only after a legitimate user has touched the fingerprint reader. Next, **the key is unlocked through user authentication**. If a legitimate user touches the sensor, the cryptographic key is unlocked, triggering a series of callback functions. Developers can **override the fingerprint callbacks** to handle different scenarios based on user legitimacy. Once authenticated, **the unlocked key can be used** by an app to encrypt, decrypt, or sign data. Google recommends using a previously generated private key to sign a server-provided authentication token to authenticate, and then to send this token to the app's remote backend [37].

4.5 Spring Security APIs

Spring Security [39] is a powerful and highly customizable framework for securing Java-based applications. It provides a wide range of security services, such as authentication, authorization, and access control. While Spring Security provides its own set of authentication features, it also supports integration with various authentication mechanisms such as *Lightweight Directory Access Protocol (LDAP)*, *OAuth*, and *Java Database Connectivity (JDBC)*. The framework enables developers to implement role-based and permission-based authorization, allowing for granular access control policies for different parts of their applications. It allows the implementation of an access-control specification model by typically defining various filters for finding access to a given resource. In addition, Spring Security offers features for securing communication between different application components using HTTPS, SSL/TLS, and other encryption mechanisms. In our review, one study [S38] examined Spring Security's use for implementing an access-control specification model.

4.6 SafetyNet Attestation APIs

Attackers can alter an application's behavior either by directly modifying the app or by obtaining root access to the host system and injecting malicious code. Hence, developers need to ensure their app's code integrity and the client device's status. Google offers the SafetyNet Attestation API [40] to check the integrity of a device or application and detect compromised devices and tampered applications. Figure 3.b illustrates the attestation process using this API. The *attest* function, requiring a *nonce* and an *API Key*, triggers the attestation API. The nonce is generated by the application's backend server and sent to the device upon attestation request (step 1). The API Key is created using Google's *API Console*, a platform for developers to manage their Google APIs. When attestation is requested (step 2), *Google Mobile Services (GMS)* conducts several checks on the device, forwarding the results to Google's server (step 3). Google's server responds with signed attestation data (step 4), which GMS delivers to the client (step 5). The client app extracts a *JSON Web Signature (JWS)* from the data, sending it to the backend server for validation (step 6), followed by the client server verifying the JWS (step 7). In our review, one study [S69] analyzed the usage of attestation API in Android applications.

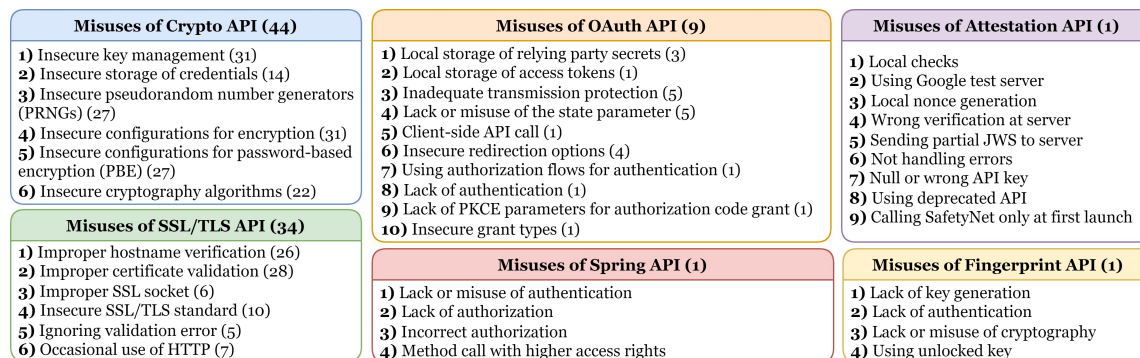


Fig. 4. List of misuses for each type of security APIs with number of studies given within parentheses

5 RQ2: MISUSES OF SECURITY APIS

This section addresses RQ2 by providing insights into various types of security APIs misuses. Through thorough analysis and semantic categorization, we have identified 39 distinct misuse types, as given in Figure 4. In the subsequent subsections, we outline these misuses focusing on each security API separately.

5.1 Misuses of cryptographic primitive APIs

The analysis of our review provided six common types of misuses made by developers while using cryptography APIs that are related to insecure management and configuration of keys, credentials, and cryptographic algorithms. Following we provide a concise description of each of the misuses.

1) Insecure key management: Inadequate key management practices pose significant security risks in cryptography. For instance, deterministic key generation and poorly derived keys lacking randomness are considered insecure. The use of hard-coded, static (constant), or predictable keys and expired or previously used keys are among the misuses identified by prior research [e.g., S1, S2, S4, S5]. Furthermore, keys of inadequate length also increase the possibility of a brute-force attack; for example, RSA keys and ECC keys should be at least 2048 and 224 bits long, respectively [41]. Another risky practice is retaining keys in memory for an extended period after their use to establish a secure channel, as it creates opportunities for code injection or side-channel attacks, enabling potential key recovery [S13]. In addition, developers must ensure that keys are distributed securely. Using a key agreement protocol that allows a single peer to generate the shared secret without involving the other peers compromises the security of negotiated key [S13]. Likewise, key exchange without authentication from a trusted entity leads to vulnerabilities to MitM attacks, where malicious servers can impersonate trusted servers and gain access to sensitive information [S15]. Additionally, developers should conduct integrity checks after symmetric key exchange to ensure the integrity of the exchanged keys [S29].

2) Insecure storage of credentials: Developers sometimes use strings to store credentials (passwords, secret keys, etc.). Programming languages, such as Java, do not clear string values from memory unless the garbage collector runs, which leaves them vulnerable to credential dumping [42]. Similarly, storing sensitive data in files or shared storage poses a significant security risk of data exposure. It is recommended to use the key storage method offered by crypto APIs for storing sensitive data such as secret keys and certificates, which requires a password to access the stored data [43]. However, developers often make sensitive data vulnerable to exposure by using hard-coded, static (constant), or predictable passwords for key storage as highlighted by several studies [e.g., S1, S2, S10].

3) Insecure PseudoRandom Number Generators: Insecure PseudoRandom Number Generators (PRNGs) are a major source of cryptography vulnerabilities [44]. It is essential to exclusively use the secure PRNG offered by crypto APIs and provide a truly randomly generated seed for initialization. However, developers often make two common mistakes: 1) using simple PRNGs that have been proven to be insecure as they can generate deterministic and predictable random numbers [45], and 2) using static (constant), low-entropy, predictable or previously-used seeds. These practices severely undermine the security of cryptographic materials, like keys, that should be generated randomly.

4) Insecure configurations for encryption: One common misuse is using unsafe *modes* of operation for encryption, such as Electronic Codebook (ECB). ECB mode encrypts data blocks independently, transforming identical message blocks into identical ciphertext blocks, thus revealing data patterns and compromising confidentiality. To ensure security, it is recommended to use more secure modes, such as Cipher Block Chaining (CBC) or Galois/Counter Mode (GCM). Some other instances of unsafe encryption modes include using DESede with ECB, DES with CBC3 SHA, AES with CBC and PKCS5Padding, CBC without HMAC, and 3DES with EDE CBC SHA. Additionally, *Initialization Vectors (IVs)* are used in several encryption modes to add entropy to ciphertexts. To ensure the security of cryptographic schemes, IVs must be randomly and properly generated. However, some developers introduce vulnerabilities by using empty, zeroed, hard-coded, static, badly-derived (e.g., deriving from keys or messages), short-length, previously-used, or any kind of predictable IVs. Another parameter that requires secure configuration is the *padding* scheme, which specifies how to fill the last block of data in encryption if its size is less than the block size. Missing padding or using insecure padding (e.g., PKCS 1-v1.5 for RSA) make it easier for an attacker to launch a padding oracle attack and recover the plaintext.

5) Insecure configurations for Password-Based Encryption (PBE): PBE is a commonly used method for generating a strong secret key based on user-supplied passwords. However, the security of PBE heavily relies on selecting appropriate parameters for key derivation, including *salt*, *password*, and *iteration count*. Improperly setting these parameters can significantly compromise the security of the derived key. One major misconfiguration of PBE is using an empty, static (constant), short-length (size < 64 bits [46]), or predictable salt, which introduces vulnerabilities to brute-force and dictionary attacks. Additionally, using hard-coded, static, weak, NIST-blacklisted, expired, previously-used or predictable password is among the other misuses identified in PBE. Moreover, developers may prefer to choose small iteration counts (less than 1000 [46]) to achieve better performance, making easier for attackers to perform brute-force attacks.

6) Insecure cryptography algorithms: Security is a constantly evolving area, and what was once considered a secure algorithm or technique may no longer be considered safe due to new vulnerabilities and attacks that are discovered over time. This makes it challenging for developers to keep up with the latest updates and current best practices in cryptography. Using unsafe symmetric encryption algorithms such as 64-bit block ciphers (e.g., DES, IDEA, Blowfish, RC4, RC2), weak password-based encryption algorithms (e.g., PBKDF1), insecure asymmetric ciphers (e.g., RSA, ECC), insecure cryptographic MACs and broken hash functions (e.g., SHA1, MD5, MD4, MD2) as well as insecure combinations of encryption and hashes or MACs (e.g., PBKDF with < SHA224) are common types of misuses identified by our review.

5.2 Misuses of SSL/TLS APIs

SSL/TLS APIs are intended to establish secure communication channels against various types of attacks, especially MitM attacks. However, misusing SSL/TLS APIs can compromise the security of the communication and leave it vulnerable to MitM attacks. Our review identified six common types of misuses made by developers while using SSL/TLS APIs:

1) Improper hostname verification: Hostname verification is a crucial security measure that ensures the hostname in the SSL certificate matches the server hostname to which the client is trying to connect. However, various studies [e.g., S1, S2, S8, S9] have revealed that some developers trust all hostnames or do not verify the hostname correctly.

Improper hostname verification enables an attacker to intercept the communication between the client and the server by presenting a valid SSL certificate for a poisoned hostname.

2) **Improper certificate validation:** Certificates serve as a means of authentication and establishing trust between the client and server, so improper certificate validation can leave the SSL channel vulnerable to MitM attack. However, many developers make mistakes in implementing proper certificate validation as identified by several studies [e.g., S14, S15, S19]. One of the most common mistakes is blindly trusting all certificates, allowing attackers to present fake certificates and gain unauthorized access to sensitive information. Additionally, some developers only check that each certificate in the chain has not expired without performing any other validation. Other ways of compromising certificate validation include incomplete validation, neglecting to check for expiration or revocation, trusting self-signed certificates, trusting too many CAs, trusting certificates with unclear names, inadequate CA verification, or insecure certificate pinning.

3) **Improper SSL socket:** The SSL socket is designed to establish a connection between a specific host and a specific port. Nonetheless, verifying and authenticating the server's hostname is essential before establishing the connection. A flawed implementation of the SSL socket may ignore hostname verification when creating the socket [S1, S26].

4) **Insecure SSL/TLS standard:** TLS, the successor of SSL, is generally considered to be more secure. However, older versions of TLS, including TLS 1.0 and TLS 1.1, have been found to be susceptible to various types of attacks, such as POODLE, BEAST, and CRIME, and therefore are no longer deemed secure. These outdated versions have been deprecated [47–49], and TLS 1.2 is being recommended as the minimum protocol version for secure communication. Nevertheless, some developers still use outdated versions of TLS and compromise the security of transmitted data.

5) **Ignoring validation error:** Some developers choose to prioritize functionality over security by ignoring errors that occur during certificate validation and instead call a method to proceed with normal operations [S24, S34, S43, etc.].

6) **Occasional use of HTTP:** Incorporating both secure and insecure connections within the same application is an unsafe practice that is occasionally adopted by some developers. This practice exposes the application to potential attacks like SSL stripping [50, 51], wherein a malicious actor can launch a MitM attack on an SSL connection.

5.3 Misuses of OAuth APIs

Generally, OAuth APIs are too complex for developers to fully understand and use correctly. The initial design of OAuth for authorization adds further challenges for developers for its adoption in authentication. Our review identified ten misuses in nine studies regarding the use of OAuth API for authentication and authorization.

1) **Local storage of RP secrets:** Although RP secrets are intended to be used by SPs to authenticate the RP, many developers store these secrets on the client-side application, posing a significant security threat to the user's privacy [S30, S48]. As a result, an attacker could retrieve the secret through reverse engineering, and use it to impersonate a benign application and request an access token from the SPs.

2) **Local storage of access tokens:** The storage of access tokens is a critical security concern as they provide RPs access to a user's protected resources hosted by SP. Simply storing them on client devices can easily compromise user accounts. Attackers may steal locally-stored tokens, consequently gaining unauthorized access to user accounts and sensitive information stored in SPs.

3) **Inadequate transmission protection:** OAuth security heavily depends on the secure transmission of messages during the OAuth process. Transmitting messages in plaintext lets attackers eavesdrop and steal access tokens or other OAuth credentials. It is crucial to use SSL/TLS encryption when transmitting confidential information during OAuth transactions to prevent this exploit.

4) Lack or misuse of the state parameter: The state parameter safeguards user sessions against a Cross-Site Request Forgery (CSRF) attacks by verifying request authenticity. In CSRF an attacker uses a user's previous session data to make a malicious request on their behalf [S52]. OAuth guidelines recommend generating and validating a randomized state parameter, bound to the user's session to prevent such attacks [22]. However, developers may misunderstand its purpose leading to mistakes such as using a predictable or constant value, enabling multiple replays, neglecting state parameter verification, accepting requests without a state parameter, or assuming that all state parameters generated by their app are valid without proper session binding checking [S51].

5) Client-side API call: A significant security concern in OAuth authentication flows arises from the reliance on client-side API calls, which attackers could easily manipulate [S30]. Some developers mistakenly assume that access tokens granted by SPs are only valid for their application. However, an attacker can use an access token granted for a malicious application to log in as a user for a different, benign application and gain access to sensitive information.

6) Insecure redirection options: To ensure the security of OAuth transactions, it is crucial to use secure methodologies for handling redirection [22]. Insecure redirection methods can allow attackers to redirect users to arbitrary domains or URLs, potentially leading to further attacks or data theft. For instance, in a mobile context, using WebView is considered insecure as it undermines the isolation between a SP and a RP [S30]. A malicious RP can use the WebView of their mobile applications to host a SP, allowing them to access the user's cookies and log in on the user's behalf.

7) Using authorization flows for authentication: OAuth was primarily designed for authorization, and its use for authentication was not explicitly defined in the initial specifications [S30]. Consequently, many developers erroneously adopted authorization flows for authentication purposes [S48]. Particularly, an access token in authorization flows is used as a means of authentication, but the access token only validates the authorization granted to a third-party application and can not provide an accurate representation of the user's identity.

8) Lack of authentication: In the OAuth transactions, a SP is responsible for authenticating a RP, and reciprocally, a RP is also responsible for authenticating a SP. This verification process can be performed using the same methods by which a SP authenticates a RP application. However, a study conducted by AuthDroid [S47] on a collection of Android apps revealed that none of the RP apps in their investigation verified the SP's identity.

9) Lack of PKCE parameters for authorization code grant: The authorization code grant is generally considered to be the most secure OAuth grant type. However, it is still susceptible to code interception attacks, where an attacker intercepts the authorization code sent by the SP and uses it to obtain an access token [52]. To mitigate this vulnerability, the Proof Key for Code Exchange (PKCE) protocol was introduced in the OAuth 2.0 specification [53]. PKCE verifies that the requesting application is the same one that originally requested the authorization code by using a cryptographically linked code verifier and code challenge exchanged between the application and the SP. PKCE is recommended as a mandatory security measure for public clients to enhance the security of the authorization code grant.

10) Insecure grant types: The security of an OAuth transaction highly depends on the choice of grant type. It is essential to avoid using insecure grant types such as implicit for authentication. Implicit grants raise a major security concern because the access token is not bound to the intended RP which enables an attacker to use a user's access token, issued to the malicious application, to log in as the user on a benign application [S48]. Best current practices recommend using the authorization code flow, which can be protected by PKCE, as a more secure alternative to other grant types [S54].

5.4 Misuses of Fingerprint APIs

Several security issues are identified based on the analysis of the usage of Fingerprint API in Android applications, with "weak usage" being the least secure [S37]. Authentication in a weak usage scenario relies solely on TEE to validate

the authenticity of a user interaction, without employing encryption. The absence of a trusted User Interface (UI) in Android allows attackers to impersonate the operating system, bypassing authentication and authorization for server transactions. Thus, root attackers can easily bypass Fingerprint API, and non-root attackers can exploit confused deputy problems³ through UI attacks. We categorized the following weak use cases as Fingerprint API misuses.

- 1) **Lack of key generation:** Some developers neglect using methods for generating keys to be used for securing fingerprint-based authentication causing insecure app development.
- 2) **Lack of authentication:** This occurs when developers designate authentication methods as null.
- 3) **Lack or misuse of cryptography:** This occurs when developers do not utilize any cryptography operation after the user touches the sensor or perform an insecure cryptography operation using constant encryption keys.
- 4) **Using unlocked key:** The key being used is not locked, and therefore, any attacker with root access can use it without requiring the user to touch the fingerprint sensor.

5.5 Misuses of Spring Security APIs

In our review, one study [S38] targeted the detection of access-control vulnerabilities arising from incorrect usage of Spring Security. This study identified four common mistakes made by developers while implementing an access-control specification model using Spring Security.

- 1) **Lack or misuse of authentication:** A common security issue is when a developer fails to implement authentication for accessing a resource, allowing unauthorized access to a resource that actually requires authentication. This can happen if the developer forgets to include an authentication filter or improperly configures a filter that grants all users unrestricted access to the resource. Consequently, any user without authentication can access the resource, resulting in a vulnerability known as missing authentication (CWE-306) [54].
- 2) **Lack of authorization:** A developer may fail to include appropriate authorization filters for a particular resource, which needs valid authorization according to the access-control specification model. Using authentication as the authorization filter also results in the same misuse since it only verifies the user's identity. The misuse causes the security vulnerability of missing authorization (CWE-862) [55], leading to either all users or authenticated users gaining access to the resource based on the type of applied filter.
- 3) **Incorrect authorization:** A developer applying an incorrect authorization formula while using an authorization filter for a specific resource causes unauthorized users to access resources. This misuse leads to the vulnerability of incorrect authorization CWE-863 [56].
- 4) **Method call with higher access rights:** This misuse is another instance of the incorrect authorization vulnerability (CWE-863). It occurs when a developer correctly configures a resource but calls a method that demands higher access rights in a deeper application layer, which should not be accessible to the user.

5.6 Misuses of SafetyNet Attestation APIs

In our review, one study [S69] conducted an analysis of how the SafetyNet Attestation API is used in Android applications. It identified various ways in which the API can be misused, including:

- 1) **Local checks:** The SafetyNet Attestation returns a JWS object representing the device and application state. It is crucial to send the JWS object to the backend server for verification. Performing local checks enables an attacker to bypass the verification by modifying the application.

³Confused deputy problems refer to security vulnerabilities that arise when a program or system mistakenly grants excessive privileges to another program or user.

2) **Using Google test server:** Google provides a verification service for SafetyNet, which is essentially a test server that allows a client application to submit a SafetyNet JWS for verification. It is important to note that this service is exclusively designed for testing purposes, and using it in a production environment may compromise the performance of the SafetyNet Attestation.

3) **Local nonce generation:** The purpose of using a nonce argument in the attest function is to avoid a replay attack. This value will be included in the JWS output of the API and can be checked against the value passed to the function to confirm that the correct JWS result is being attested. However, if the nonce value is generated locally on a compromised device or application, an attacker can exploit a previously generated nonce value to conduct a replay attack.

4) **Wrong verification at server:** The verification process of SafetyNet JWS involves several checks by the server, such as validating the nonce, APK package name, and the hash of the application's signing certificates present in the JWS payload. Inaccurate or incomplete execution of these validations may enable an attacker to send a tampered SafetyNet JWS to the server and bypass the verification.

5) **Sending partial JWS to server:** The SafetyNet JWS should be sent to the server for verification. However, a developer may choose to send only certain values extracted from the JWS object. This enables attackers to replace the missing values on a compromised device or application without any means for servers to detect tampering.

6) **Not handling errors:** Errors may occur during the integrity checks performed by SafetyNet Attestation. To ensure a successful attestation process, handling any errors that arise is important. This can involve attempting the attestation again or following the protocol if integrity checks fail.

7) **Null or wrong API key:** Developers must provide the API with a valid key obtained from the Google APIs Console. However, it is not uncommon for developers to mistakenly use an incorrect or null API key, leading to an error in the attestation process. If this error is not handled properly, the attestation process fails, leaving any tampering undetected.

8) **Using deprecated API:** The attestation process cannot be accomplished if developers use the deprecated API, which always returns an error and can not generate a valid SafetyNet JWS.

9) **Calling SafetyNet only at first launch:** SafetyNet Attestation should be consistently performed during an application life cycle, specifically when launching or handling sensitive information. However, some developers only perform SafetyNet Attestation during the first launch, leaving the application vulnerable to tampering. Attackers can launch the application once in a non-tampered state, then tamper the device or application later without being detected, as SafetyNet Attestation will not be performed anymore.

6 RQ3: MISUSE DETECTION TECHNIQUES

To ensure proper usage of APIs, one needs to follow the specifications provided for input, output, and invocation context. These specifications evolve over time and detecting misuses involves verifying if the application adheres to the most up-to-date specifications available. *Heuristic-based* approaches are the most commonly used method to detect security API misuse. They use API specifications to model normal and misuse patterns of API use and then compare the application being tested with those patterns to identify misuse. They can also generalize code examples, including secure or insecure use of APIs to infer patterns for detecting misuse. *Machine learning (ML)* techniques have recently emerged as an alternative approach, which learns a misuse detection model directly from source code available in code repositories. This approach requires the extraction of features representative of correct or incorrect API use, and then a model is trained to classify input code as secure or insecure.

One way to assess an application's security API usage is by analyzing its *source code*, which provides early feedback to developers in the development phase. However, a tester may execute an application and examine *runtime information*

in order to detect misuses, without the need for access to source code. Figure 5.a provides an overview of techniques used to detect misuses of security APIs with three main components, **input**, **analysis engine**, and **output**. Figure 5.b illustrates our proposed taxonomy, which categorizes current literature based on modeling input, testing input, output types, automation mode, and analysis algorithms. The following subsections present an analysis of the misuse detection approaches based on each factor.

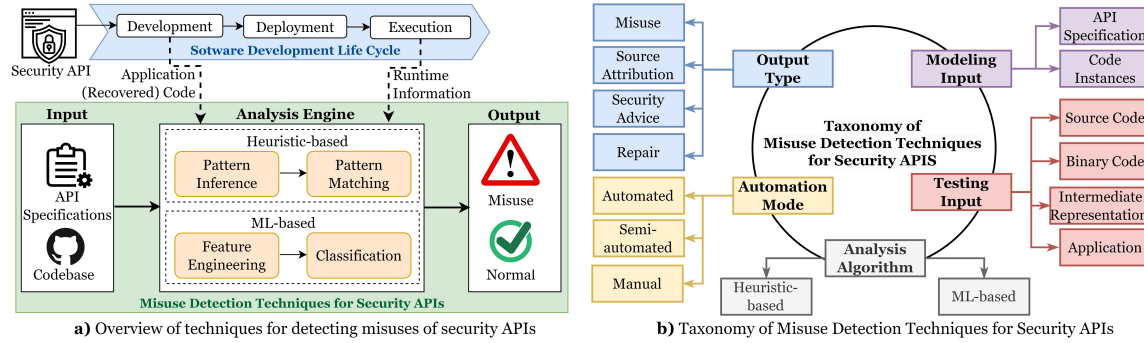


Fig. 5. Overview and taxonomy of techniques used by primary studies for detecting misuses of security APIs

6.1 Modeling Input Types

The analysis engine can use different input types such as API specifications and code examples to build a misuse detection model. A primary source to distinguish between misuse and normal usage of APIs is **API specifications** that are typically published upon API release. These specifications can be obtained from various sources such as API documentation, the literature related to the vulnerabilities caused by misuse of security APIs, or security-related standards and the best current practices defined by organizations and agencies such as NIST and IETF [S10]. In total, 63 primary studies used API specifications to manually define misuse or normal patterns for security APIs and then used a pattern-matching component to identify misuses. However, API specifications may become obsolete over time and may not reflect current practices for the correct use of APIs. For example, using the SHA-1 hash function is no longer recommended for cryptography since the discovery of a collision against it [57]. Hence, relying solely on manually defined patterns based on these specifications may lead to failure to detect new or evolving misuses.

Another valuable source for modeling is **code examples**, which have been adopted by 3 studies to infer misuse and normal patterns for crypto APIs [S2, S29, S46] and SSL APIs [S2, S46]. Code repositories are frequently updated to fix newly discovered security issues. Hence, CryptoChecker [S29] introduced DiffCode to infer crypto misuse patterns from code changes, including three steps: (i) collecting code changes from GitHub repositories, with a focus on patches for classes that use the target API; (ii) creating Directed Acyclic Graphs (DAGs) to represent the invoked APIs and related parameter values, and comparing them to extract relevant changes; (iii) clustering similar changes to identify API misuse patterns. On the other hand, Seader [S2] and VuRLE [S46] leveraged Abstract Syntax Trees (ASTs) of a given vulnerable code and its corresponding repaired code to infer misuse patterns from code examples. Unlike heuristic-based approaches that consist of pattern inference and pattern matching, ML-based approaches learn detection models directly from source codes available in code repositories. Three studies [S18, S45, S65] adopted a supervised learning approach to learn a detection model from secure and insecure code examples with API usage. The obtained model is then used to predict an API use class as misused or normal, for an application under test. Only 6 studies among the 69 rely on code examples for API misuse detection whereas others rely on API specification as shown in Figure 6.a.

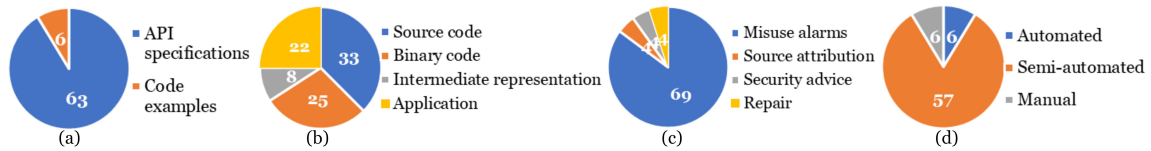


Fig. 6. Distribution of misuse detection techniques adopted by primary studies over (a) modeling input types (b) testing input types (c) output types (d) automation mode

6.2 Testing Input Types

Various input types are utilized to analyze an application for the presence of misuse. Out of the 69 studies reviewed, 33 examined the *source code* of applications for detecting instances of security API misuses. While their scope is restricted to open-source applications, they offer developers timely feedback on potential misuse. Some studies proposed misuse detection tools as an IDE plugin, providing real-time feedback to developers. For example, FixDroid [S5] is a plugin for Android Studio, which highlights bad security practices, including misuse of security APIs as developers write code. Similarly, CogniCrypt [S3] was introduced as an Eclipse plugin, which uses standard error markers by Eclipse to issue misuse alarms. Some studies [S14, S23, S26] decompiled binary codes to gain access to source codes when the original source code is unavailable. However, obfuscation techniques that modify class files to protect the source code present challenges to successful decompilation. The proposed approaches targeting source code analysis are limited to specific programming languages such as Java, Python, C, or C++. The need for multi-language analysis tools was highlighted in one study [S20], which revealed that developers in the embedded domain frequently rely on C language to implement cryptography in MicroPython projects.

Furthermore, some studies perform program analysis on *binary codes* in the absence of source code. Binary files are typically packaged with other resources such as images and configuration files into application files. Thus, misuse detection techniques usually perform *reverse engineering* to decode an APK file and disassemble it into binary files, ready for analysis. Disassembling is typically faster than decompiling and is not affected by obfuscation. Common program analysis frameworks used by primary studies to extract binary files from application files include SOOT [58], APKTool [59], and Androguard [60]. In total, 25 studies targeted analyzing binary codes to detect misuses in a variety of applications such as Android [e.g., S1, S2, S4, S7, S11] or Apache projects [S2, S11].

However, analyzing low-level representations of applications (i.e., binary files) can be laborious and error-prone. Some studies [S4, S11, S12, S28, S37, S44] chose to convert binary files into a higher-level *Intermediate Representation (IR)* that accurately captures the program behavior and is more appropriate for analysis. In addition, using IR helps to address the inconsistency of different input formats. For instance, CryptoREX [S22] converts diverse binary codes into a unified IR to carry out large-scale crypto misuse detection for IoT firmware with non-unified underlying architectures. Moreover, the research conducted in [S45] performed misuse detection by identifying instances of specific vulnerable code snippets with security API misuse in Android apps. This study used WALA [61]—a program analysis framework—to transform both known vulnerable code and Android apps to a unique IR.

Meanwhile, some approaches rely on *application files* to detect misuses by executing them and observing their behavior during runtime, without needing to access source or binary code. However, they are resource-intensive and time-consuming, requiring the deployment and execution of an application. In our review, 22 studies [e.g., S6, S8, S10, S13, S14] used application files as input for testing their analysis engine.

Figure 6.b shows the distribution of input types used for testing software artifacts in our reviewed studies. It is worth noting that some studies conducted the analysis using multiple input types. We have listed and counted them under

their respective relevant categories. For example, some studies [e.g., S19, S24] detect potential misuse by analyzing binary codes and then execute the applications to identify exploitable misuses during runtime.

6.3 Output Types

Common ways to support developers to deal with misuses are issuing misuse alarms, performing source attribution, generating security advice, and assisting with fixing misuses. Figure 6.c shows the distribution of output types in our reviewed studies. All the studies that provided support, aside from misuse alarm, are focused on crypto APIs.

Source attribution: The objective of source attribution is to attribute whether a misuse originates from an application code or from a third-party library. It is beneficial for developers to identify libraries with misuses and avoid using them, and for researchers to avoid over-counting misuses by identifying those that stem from libraries [S12]. Three studies investigated the primary source of crypto misuse and, interestingly, showed that third-party libraries are the major reason for misuse in Android applications [S1, S12] and popular Python projects in GitHub [S20]. Thus, it is crucial for developers to be aware of the security implications of the libraries. Furthermore, one study [S35] evaluated a large database of third-party libraries and found that crypto misuses are very common among widely used advertising libraries. The study also identified affected Android apps through third-party library detection.

Security advice: Developers often face challenges while trying to fix identified issues, leading to new mistakes [62]. Warning messages can support developer in fixing misuses and writing more secure code by providing them with clear and helpful feedback. According to one study [S59], integrating security advice and secure programming hints with crypto APIs led to significant improvements in the accuracy and security of API usage in users' codes. It designed a patch that hooks crypto API calls to identify insecure cryptographic objects. When a misuse is detected, it triggers an advice method that utilizes contextual information to issue a warning message. Similarly, FixDroid [S5] was developed as a plugin for Android Studio, which provides real-time feedback and suggestions for quick fixes. Besides, studies [S63, S17] focused on tutoring and helping students to correctly use crypto APIs. One study [S63] implemented an IDE-based interactive educational environment as an intelligent tutoring system, to help students fix misuses. Another study [S17] designed CryptoTutor for providing coding feedback to help students understand the misuse detected in their program.

Repair: Studies that generate security advice and provide users with fix suggestions, generally lack the ability to provide customized fixes for a given vulnerable program. Only four studies [S2, S16, S17, S46] proposed the automated generation of customized fixes for crypto misuses. One simple approach is to use manually crafted patch templates, as employed by CDRep [S16] and CryptoTutor [S17]. These studies provided a set of templates consisting of code transformations that were applied to vulnerable programs. However, such patch templates are unable to cover all possible variations of misuse. In contrast, Seader [S2] and VuRLE [S46] used automated techniques to generate customized fixes from insecure and secure code examples. They generated and compared ASTs of a given vulnerable code and its corresponding repaired code to extract the edit operations necessary to transform a vulnerable program into a secure one. Using edit operations, Seader generated an abstract fix for a vulnerable program including abstract variables which were replaced by concrete variables to customize fixes. On the other hand, VuRLE customized the edit patterns for a specific vulnerable program and applied the customized changes to repair vulnerabilities.

6.4 Automation Mode

The reviewed studies are categorized into *manual*, *semi-automated*, or *automated* groups based on their level of automation. Figure 6.d shows the distribution of analysis techniques with regard to their level of automation. Most of the reviewed studies (57 studies) fall under the semi-automated group. This group relies on manually defined patterns

for correct or incorrect patterns of API usage. Further details on semi-automated approaches are provided in Section 6.5. Six studies [S2, S18, S29, S45, S46, S65] automatically learn detection models from code examples representing correct and incorrect use of security APIs. Although this approach eliminates the need for manual effort, it requires labeled datasets to train the detection models. Six studies [S35, S48, S60, S61, S64, S66] detected misuses through manual source code inspection by analyzing the state of security API use in real-world software artifacts such as GitHub open-source projects [S61] or code snippets found in developers' forums posts [S64, S66].

6.5 Analysis Algorithms

We have categorized the analysis algorithms of reviewed studies into two high-level groups: *heuristic-based* and *ML-based*. In our review, 66 studies applied a heuristic-based algorithm, whereas only three studies employed ML-based algorithms. This section provides insight into the approaches adopted in each category.

6.5.1 Heuristic-based. Heuristic-based algorithms have been widely used to identify misuse of security APIs in software applications. These algorithms typically involve modeling patterns for correct or incorrect usage of security APIs and then applying program analysis techniques to identify whether the application being tested matches these patterns. Figure 7 shows the taxonomy of heuristic-based approaches based on the adopted pattern types, pattern representation models, and program analysis techniques, which are elaborated on below.

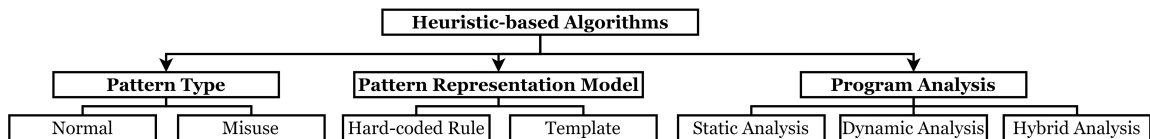


Fig. 7. Taxonomy of Heuristic-based algorithms for detecting misuse of security APIs

A. Pattern Types: API misuses can be detected using two types of patterns: *normal* or *misuse*. Misuse detection through normal usage patterns involves modeling correct API use and identifying deviations from these patterns in a given application as misuse. This approach has been employed in 12 studies [S3, S9, S11, S36, S40, S51, S53-S56, S58, S61], and can identify a wide range of misuse types through normal patterns that are limited in scope. While APIs can be misused in numerous ways, only a small subset corresponds to proper usage. However, a drawback of this approach is that it produces high false alarms when it fails to model normal patterns thoroughly, leading to unmodeled normal patterns being mistakenly classified as misuses. An alternative solution is to model incorrect API uses and identify matches of an application with these patterns as misuses. However, predicting all possible ways that a developer could misuse an API is a challenging task, so this approach may not capture all misuses. Nevertheless, most of the studies in our review (54/66 studies) rely on misuse patterns to avoid high false alarms in approaches based on normal patterns.

B. Pattern Representation Model: The simplest way to model patterns is to establish a fixed set of rules that will be *hard-coded* in the misuse detection algorithm. The consistency of an application is then evaluated against these rules. Some example rules that were used to model misuse patterns are “Don’t use a constant key for encryption” while using crypto APIs, or “Don’t store access tokens on clients” while using OAuth APIs. As illustrated in Figure 8.a, most of the primary studies perform misuse detection based on hard-coded misuse patterns. For example, CryptoLint [S4] hard-coded six misuse patterns for detecting misuses of crypto APIs, which were later used and expanded in other studies [e.g., S12, S16, S29, S35]. 31 studies use hard-coded misuse patterns to detect misuses of crypto APIs. Hard-coded

rules were also used to model misuses of SSL/ TLS [e.g., S8, S14, S19], OAuth [e.g., S47-S50], Fingerprint [S37], SafetyNet Attestation [S69] APIs. Furthermore, one study [S54] based on hard-coded normal patterns investigated the compliance of Android apps with the best current practices of OAuth for native apps.

Hard-coded rules-based approaches are straightforward to design and implement but have limitations. For instance, they can only detect a predefined set of misuses, making it difficult to extend beyond these established rules. The constantly evolving threat landscape for security APIs requires developing more adaptable methods for pattern representation models. With this goal, some researchers proposed using templates to abstractly represent both secure or insecure uses of security APIs that include language-based, graph-based, state-machine, and code-based templates.

Language-based templates rely on a syntax-based representation of patterns. CrySL [S11] is a language designed for crypto experts to specify the secure usage of crypto APIs. Several studies [S3, S11, S36, S40, S58, S61] used CrySL to detect crypto API misuses. Meta-CrySL [S55] is an extension of CrySL that helps manage variations in the API and security standards specified in CrySL. Another study [S56] introduced a formal model for security annotations that describe properties ensuring the secure usage of WebCrypto APIs within a JavaScript program. Furthermore, an anti-protocol language was introduced by the study [S30] to describe common misuse patterns for OAuth API.

Graph-based templates involve nodes that represent key elements while using APIs and edges that represent correlations between these elements. SSLint [S9] models the proper use of SSL API based on the program dependency graph representing critical API call-sites, variables, parameters and conditions.

Finite State Machine (FSM) can represent the behaviour of an application while using an API through a finite number of states and transitions between them. For example, two studies [S51, S53] used FSMs to model the regular operation of OAuth, where sending an HTTP(S) request or receiving an HTTP(S) response triggers the transition between states. FSMs were also used to model misuse patterns of SSL/TLS [S25] and Spring [S38] APIs. The research conducted in [S38] implemented FSMs to monitor the program's authorization state for each type of misuse. The transitions between states occur when method calls are made to authorize the user or gain access to a critical resource.

Code template: Crafting language, graph, and FSMs-based templates relies on domain knowledge and manual effort to specify the critical elements for API usage, their correlation, and modeling them as templates. Source code is an alternative solution that includes instances of security APIs used to automatically infer templates useful for misuse detection based on code analysis. A misuse template abstractly represents a code pattern including misuse of security APIs. Two heuristic-based studies [S2, S46] relied on concrete (insecure, secure) code examples to generate misuse and repair templates. The study [S2] extracted the edit operations by comparing a given vulnerable code's ASTs and its corresponding repaired code. Then, vulnerable code templates and repair templates were generated by performing a data-dependency analysis of ASTs and abstracting variables in the code. The vulnerable code template is used to detect misuses through pattern matching, while the repair template is used to generate customized fixes. The study [S46] similarly used ASTs to extract the required edit operations for fixes and then clustered similar edit operations based on the longest common subsequences between them. Finally, each cluster was generalized to a vulnerable code template and repair template to be used for detecting and fixing misuses. While code templates facilitate automatic pattern generation, they are limited to known misuses for which code is already available. In our review, one study [S2] used a set of code examples from prior research and another study [S46] collected 48 applications from GitHub and identified misuses by manual analysis of commits within each application's repository. Table 5 summarizes descriptions, strengths and weaknesses of pattern types and representation models and Figure 8.a shows their distribution in primary studies.

C. Program Analysis Techniques: Our review identified three categories for dividing program analysis techniques

Table 5. Pattern Inference categorizations with their descriptions, strengths and weaknesses

Type	Description	Strengths	Weaknesses
Pattern Type			
Normal	Patterns are inferred from correct uses of APIs and any violation of these patterns is considered as misuse.	<ul style="list-style-type: none"> Limited number of patterns Incomplete specification does not result in missed vulnerabilities 	<ul style="list-style-type: none"> Susceptible to high false alarm rates
Misuse	Patterns are inferred from incorrect uses of APIs and any matches with these patterns are considered as misuses.	<ul style="list-style-type: none"> Incomplete specification does not result in false alarms 	<ul style="list-style-type: none"> Difficult to capture all possible patterns Incomplete specification results in misuses being missed
Pattern Representation Model			
Hard-coded rules	Patterns are defined as a set of rules.	<ul style="list-style-type: none"> Simple to design 	<ul style="list-style-type: none"> Dependent on domain knowledge Hard to extend to new misuses
Template	Patterns are abstracted via a higher level template.	<ul style="list-style-type: none"> Easier to extend to new misuses 	<ul style="list-style-type: none"> Difficult to design a template from instances

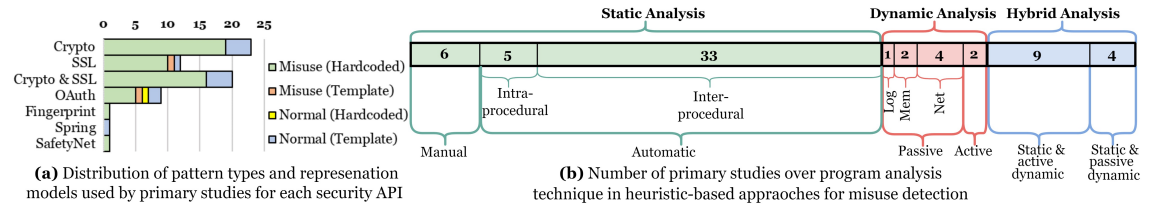


Fig. 8. Distribution of a) pattern types and representation models b) program analysis techniques in heuristic-based misuse detection

based on their reliance on code execution: **(i) static analysis**, **(ii) dynamic analysis** and **(iii) hybrid analysis**. Following we investigate these methods and their adoption in the existing literature.

(i) Static Analysis: Static analysis involves examining (recovered) source code, binary code, or an intermediate representation of binary code without executing the application. It is also known as *white-box* testing as it requires application code or its implementation details to identify misuses. It is resource and time efficient and can achieve high code coverage. The main idea is to determine the possible values of the parameter objects in a relevant API call and examine them against normal or misuse patterns to detect misuse which is achieved through **data flow analysis**. Data flow analysis typically uses *program dependency graphs* to understand how data is used and manipulated within a program. There are two types of data flow analysis: **intra-procedural** and **inter-procedural**, depending on whether the interactions between different procedures or functions are considered or not. Most of the static approaches in our review rely on inter-procedural analysis, which enables the capture of more complex misuses. However, several studies [e.g., S28, S38, S59] are based on intra-procedural analysis. For example, one study [S28] used data flow in cryptographic functions to identify paths taken by a parameter from its initial origin to its ultimate use within a function.

To achieve more efficient misuse detection, many studies applied **program slicing**. Program slicing simplifies the complexity of a program by removing parts of the code that are irrelevant to a specific analysis or task [63]. This is accomplished by computing a set of program statements that affect (backward slicing) or are affected by (forward slicing) a given slicing criterion, which is typically an API parameter, based on data flow [S1]. For instance, CryptoTutor [S17] applied inter-procedural data flow analysis and program slicing to detect crypto misuses in Java code. CryptoLint [S4] also used inter-procedural backward slicing to track flows between crypto parameters and operations, enabling the detection of pre-defined misuse patterns in Android applications. Later, BinSight [S12] and CDRep [S16] leveraged CryptoLint to examine the current state of crypto API usage in Android applications, with additional efforts towards source attribution [S12] and repair [S16]. Amandroid [S44] also applied inter-procedural data flow analysis to assess the security state of Android apps in terms of data leaks, data injection, and improper use of crypto APIs. CogniCrypt_{SAST} [S11], a tool based on inter-procedural analysis, was designed as a compiler for CrySL (a

language-based template for normal use of crypto APIs; detailed in pattern representation models) to check Java source code for compliance with CrySL and generate code for common crypto tasks. Several studies [e.g., S36, S40, S58, S61] used CogniCrypts_{AST} for detecting misuses of crypto APIs.

Several studies [S30, S48, S54] used static analysis to detect misuses in OAuth API. For example, OAuthLint [S30] used Flowdroid [64] to perform inter-procedural data flow analysis and identify key elements for misuse patterns. Another study [S37] classified applications into different security levels based on their use of the Fingerprint API, performing inter-procedural backward slicing to extract API parameters as features in a rule-based classification.

A set of studies utilized ASTs to conduct data flow analysis on source code. For instance, in one study [S20], the source code was parsed into AST, and subsequently, backward slices were generated by filtering the AST according to the targeted crypto elements. Other studies [e.g., S33, S63] also employed ASTs for data flow analysis purposes. They analyzed ASTs to identify misuse locations that match predefined patterns.

Although program slicing improves the efficiency of static analysis, it may lead to large memory and runtime overhead on massive-sized projects. To address this challenge, CryptoGuard [S1] proposed a trade-off between accuracy and scalability by performing on-demand slicing. This approach limits the analysis to methods that have the potential for security impact, effectively reducing the size of the code that needs to be analyzed. Additionally, it utilized refinement algorithms to remove irrelevant language-specific elements and mitigate the high rate of false alarms in static analysis. Later, another study [S39] used CryptoGuard to evaluate the state of crypto API use in Android applications.

Another technique to minimize false alarms in static analysis is *symbolic execution* [65] that executes a program by using symbolic values as inputs, rather than concrete values, and expressing the values of program variables as symbolic expressions of these inputs. Several studies, such as SSLDoc [S21] and TaintCrypt [S25], leveraged symbolic execution to statically detect security API misuses by creating program path traces that capture semantic information for each targeted API. Another study [S18] performed a simple variant of symbolic execution to extract crypto API sequences from Android applications, which were then used to learn probabilistic models to predict misuses. Some studies performed manual code analysis to detect crypto misuses [S35, S60, S61, S64, S66] and OAuth misuses [S48].

Developers can use static analysis tools in their daily coding tasks to detect misuses in the early stages of software development. However, the static analysis also has limitations, such as high false alarm rates caused by infeasible misuses (that never occur at runtime) and failure to capture runtime misuses.

(ii) Dynamic Analysis: Dynamic analysis involves executing the code of an application and monitoring its behavior during runtime. As a result, these approaches do not usually produce false positives and can capture misuses occurring during runtime. Dynamic analysis is also known as *black-box* testing as it treats applications as black boxes and only considers the external behavior of an application at runtime. There are two types of dynamic analysis for discovering software vulnerabilities, including API misuses: *active* and *passive*. *Active dynamic analysis* involves intentionally attempting to exploit vulnerabilities or cause disruptions in a system. In contrast, *passive dynamic analysis* focuses on collecting data and observing behavior without trying to cause harm. The passive dynamic analysis examines execution logs, the memory state of a program, or network traces to gain insight into its behavior. Considering the observed data, we have categorized passive dynamic approaches into *log*, *memory*, and *network analysis*.

Log analysis involves collecting runtime information and execution traces and performing offline analysis after the execution is completed. While the offline analysis does not affect the application's performance [S10], it can generate large log files, creating an I/O bottleneck slowdown [S13]. For example, one study [S10] examined logs that record parameters relevant to crypto API calls to find matches with some misuse patterns. In our review, log analysis was performed by a few studies to detect misuse of security APIs. *Memory analysis* was also utilized by some studies for

Table 6. Program analysis techniques with their descriptions, strengths, and weaknesses

Type	Description	Strengths	Weaknesses
Static	Static analysis examines the application's code against API usage constraints.	<ul style="list-style-type: none"> Doesn't require program execution and is scalable to a large number of applications. 	<ul style="list-style-type: none"> Applicable only to open-source projects Suffers from high false alarm rate
Dynamic	Dynamic testing executes the software and validates output or runtime information against API usage constraints.	<ul style="list-style-type: none"> Able to capture misuses occurring during runtime 	<ul style="list-style-type: none"> Requires code execution Costly and not scalable to a large number of projects

misuse detection. For example, K-Hunt [S13] tracks memory buffers that store encryption keys to verify whether keys are generated and transmitted securely. It started with a lightweight dynamic analysis to gather runtime information required to locate memory buffers where crypto keys were stored. Meanwhile, several studies adopted *network analysis* approach to detect misuses of APIs such as SSL/TLS and OAuth. One study [S52] evaluated the implementation of CSRF protection in OAuth transactions by checking the presence or absence of a state variable in URLs.

Some studies performed active dynamic analysis to verify the results obtained from network analysis. Active dynamic analysis can include a range of techniques, such as *penetration testing* which simulates a real-world attack on a running application to identify any misuses that could be exploited. It is considered as the most effective approach to uncover exploitable misuses and avoid false alarms. One study [S50] manually analyzed the HTTP messages to capture the information flow of SSO credentials and detect potential misuse of OAuth. It further designed exploits to prevent manual inspection errors. Similarly, another study [S49] performed network analysis followed by examining the feasibility of a CSRF attack to uncover exploitable misuses of OAuth.

However, dynamic analysis is resource-intensive and time-consuming, involving tasks such as installing, configuring, and testing, some of which may require human intervention [S23]. It also has limitations in terms of code coverage.

(iii) Hybrid Analysis: Attempts have been made to combine static and dynamic analysis in a *hybrid* approach to leverage the strengths of both techniques and overcome their weaknesses. Table 6 provides a concise summary of the strengths and weaknesses associated with static and dynamic analysis techniques. To mitigate the risk of false positives in the static analysis, some researchers proposed a hybrid approach that typically applies static analysis to identify potential misuses, followed by dynamic analysis to validate the results. Several studies [S8, S14, S19, S24, S57, S67] evaluated Android apps against a MitM attack to verify misuses reported by static analysis. Another study [S23] applied manual static analysis to find potential crypto misuses in Android apps and then performed dynamic memory analysis to examine the crypto libraries invoked during execution. This approach enables the detection of misuses that are feasible at runtime. Another study [S69] used a combination of static and dynamic analysis to find Android applications that call the SafetyNet Attestation API during their execution. Next, it did a manual static analysis to find vulnerable applications with potential misuses, followed by bypassing the SafetyNet Attestation checks to confirm the misuses.

Static analysis can also serve as a guide for dynamic analysis, reducing the time and memory consumption of dynamic analysis by pruning its exploration space. Some studies [e.g., S19, S24, S67] employed a preliminary static analysis to detect misuses. They further used static analysis for method call graphs to identify the entry points that trigger the execution of vulnerable methods. These entry points were then used to generate inputs for running applications during dynamic analysis, resulting in a more efficient analysis with a reduced input space. Another study [S6] combined static and dynamic analysis techniques to detect crypto misuses in iOS apps. It first used static analysis to find the locations of crypto APIs and then monitored those API calls at runtime using *API hooking* techniques. Misuses were detected by analyzing the execution logs, which record parameter values and other relevant information. AuthDroid [S47] also adopted a hybrid approach to detect OAuth misuse in Android apps. It uses static analysis to extract the basic elements of OAuth (e.g., user-agent, the identity of SP) from the app, then uses a MitM proxy in dynamic analysis to find API

misuses in an authentication process. While the mentioned studies followed a static-dynamic approach in detecting misuses, one study [S26] has taken a different hybrid approach by first simulating a MitM attack to find vulnerable apps, and then manually performing static code analysis to identify the root causes of misuses.

6.5.2 ML-based. Based on our review, only three studies adopted ML-based algorithms to detect security API misuses. The basic idea is to classify API usage instances within a given application as correct or incorrect using features that indicate the application's behavior. Following we examine the feature engineering and classification components of these approaches.

A. Feature Engineering: Three types of features were identified in the existing literature for building security API misuse detection models which are sequential-, word-, and graph-based features.

(i) Sequential-based features: API sequences representing both API orders and API arguments were used to learn probabilistic models proposed in [S18]. To this end, they used static analysis to extract possible traces for each reachable method from application binary files. Furthermore, they performed a simple variant of symbolic execution on each trace and then filtered traces of irrelevant APIs.

(ii) Word-based features: Term frequency-inverse document frequency (tf-idf) is a common technique used in Natural Language Processing (NLP) to evaluate the importance of a term in a document or corpus. Recent advances in NLP have inspired many researchers to apply it to analyzing source code by considering it as natural-language text. In our review, one study [S45] extracted tf-idf from source codes to train a misuse detection model.

(iii) Graph-based features: One study [S65] utilized graph-based features to analyze the usage of security APIs in source code. First, the source code was parsed to AST and then modeled through graph embedding techniques, Bag of Graphs (BoG), and node2vec. These techniques are similar to word embedding techniques in NLP, where words are embedded in a vector space based on their co-occurrence with other words in a text corpus. For example, BoG generates a collection of graph bag items representing elements or sub-graphs within a graph. These items are then used to construct a vector representation that captures the local attributes and relationships of the original graph. Node2vec is another graph embedding technique that extracts features from graphs, utilizing a flexible neighborhood sampling strategy.

B. Classification: Our review identified various classification techniques that were employed to detect security API misuse. In one study [S18], two probabilistic models, Hidden Markov Model (HMM) and n-gram, were trained using both secure and insecure API sequences. These models were employed to predict the probability of a given API sequence being secure. An API sequence was considered insecure if its probability fell below a pre-defined threshold. The study also addressed the problem of identifying misuse locations within an insecure sequence by using a distance measure based on the probability of an API misuse at possible locations. In another study [S45], code snippets with the usage of security APIs were mined from SO and then classified using a Support Vector Machine (SVM) model. A small set of extracted code snippets was manually labeled to build the training dataset. Similarly, the approach proposed in [S65] used SVM but trained a classifier for each category of misuse using correct and incorrect instances for corresponding misuse. Thus, the model can identify both the presence and type of API misuse.

7 RQ4: EVALUATION METHODS AND RESULTS

This section provides an insight into the prevalence of security API misuses in software artifacts by presenting the misuses reported in primary studies (Section 7.1). We also analyze metrics, benchmarks, and strategies adopted to evaluate the performance of misuse detection approaches (Section 7.2).

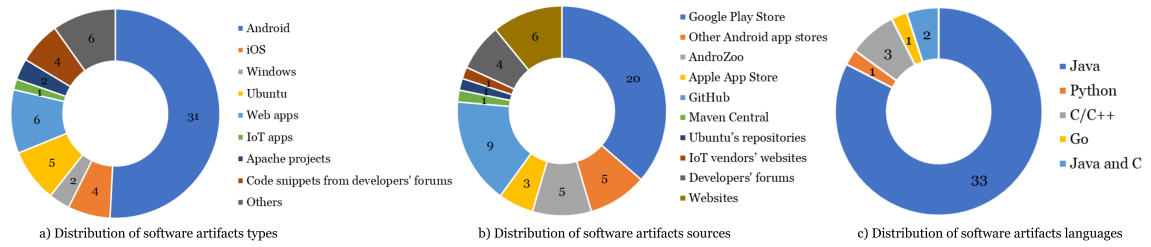


Fig. 9. Distribution of (a) types, (b) sources, and (c) programming languages of software artifacts analyzed by reviewed studies

7.1 Security API Misuses in Wild

A total of 50 primary studies have investigated the usage of security APIs in real-world projects and reported alarming results. The following subsections present the software artifacts studied and the types of reported misuses.

7.1.1 Software artifacts. Our reviewed studies designed experiments to assess various software artifacts acquired from various data sources, as demonstrated in Figure 9 (a-b). Given the widespread use of mobile devices for storing confidential data, and the fact that Android is the most widely used system, many researchers have focused on assessing the security of Android apps in terms of their use of security APIs. Most of the studies used the Google Play Store as the primary source for downloading Android applications. Moreover, AndroZoo [66], which offers the most extensive publicly available dataset of Android apps collected from the Google Play Store, was utilized in several studies. The number of apps targeted in these studies ranged from 45 [S15] to over 500,000 [S36]. Several studies focused on analyzing iOS apps from the official Apple App Store. Two studies leveraged Maven Central and GitHub to analyze Apache projects, and one study [S22] assessed 521 IoT firmware obtained from various IoT vendors’ websites. Additionally, there were studies dedicated to assessing Windows and Ubuntu applications, and studies that investigated OAuth API usage in web applications.

Previous research revealed that forums like SO are crucial sources of information for developers as they may copy and paste code snippets from such sources into their software projects [67]. Consequently, four studies examined the state of security API usage in code snippets available in forum posts. Three studies assessed 187 [S33], 3,834 [S45], and 25,855 [S66] code snippets from SO, while another study [S64] investigated 140, 71, and 48 posts from Oracle Java Cryptography (OJC), Google Android Developers (GAD), and Google Android Security Discussions (GASD), respectively. The distribution of programming languages of studied artifacts is depicted in Figure 9.c which shows the majority of the software artifacts were written in Java, followed by C/C++ and Python.

7.1.2 Reported Misuses. Our review has revealed a concerning trend of incorrect usage of security APIs among developers, with reported vulnerable applications ranging up to 100% of the studied applications in some cases.

The study [S45] identified 30.28% of code snippets from a collection of security-related snippets on SO as being insecure while applying ML techniques, with crypto and SSL/TLS misuses. It further analyzed a dataset of Android applications and found that 15.4% of this dataset contained security-related code snippets from SO, out of which 97.9% contained at least one insecure code snippet. Other studies in our review relied on heuristic-based approaches to report instances of security API misuse in real-world software artifacts.

Three boxplots, shown in Figures 10 (a-c), depict the percentage of software artifacts with at least one misuse identified by heuristic-based approaches. The boxplots represent three different approaches: static, passive dynamic, and active dynamic. Static approaches may yield false positives. Passive dynamic approaches can detect misuses during runtime, and active dynamic approaches can identify exploitable misuses, resulting in different misuse rates. Static

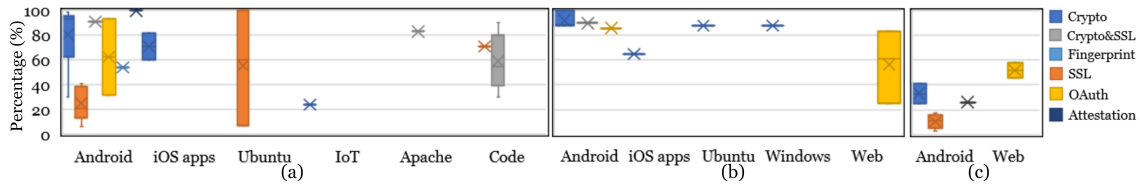


Fig. 10. Boxplots with mean markers illustrating the percentage of software artifacts with at least one misuse detected by (a) static analysis, (b) passive dynamic analysis, and (c) active dynamic analysis approaches in reviewed studies

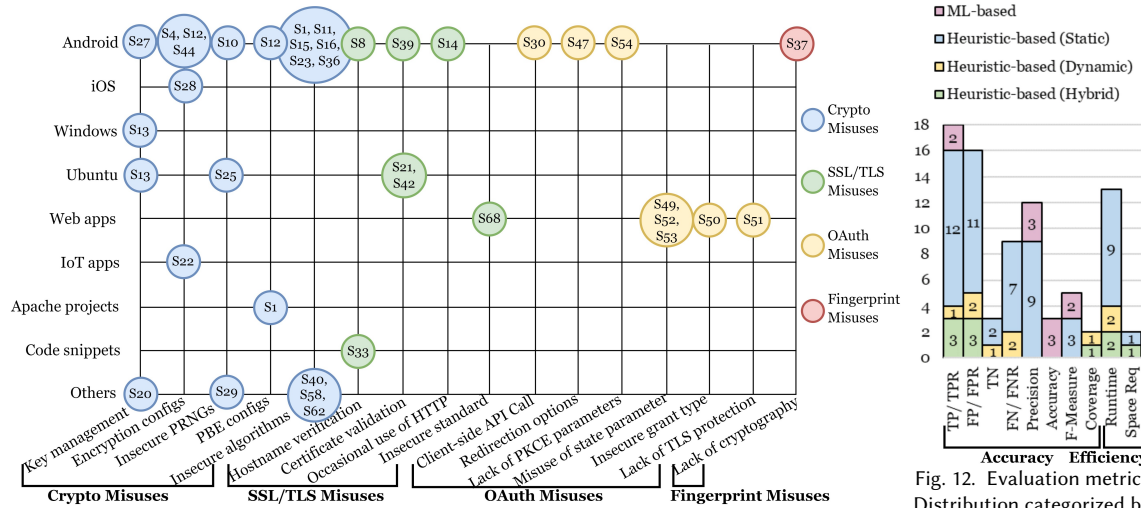


Fig. 11. Most common misuses of security APIs in real-world software identified by primary studies

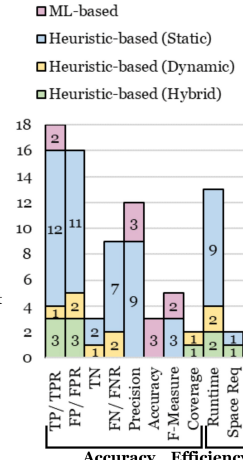


Fig. 12. Evaluation metrics Distribution categorized by detection techniques

approaches have reported at least one misuse of security APIs, including crypto, SSL/TLS, OAuth, Fingerprint, and Attestation APIs in 6%-100% of Android applications, 60%-82% of iOS applications, 7%-100% of Ubuntu, 24% of IoT applications, 83% of Apache projects, and 48%-90% of code snippets in developers' forums. Passive dynamic approaches have reported at least one misuse of security APIs, including crypto, SSL/TLS, and OAuth APIs in 86%-100% of Android applications, 65% of iOS applications, 88% of Ubuntu and Windows applications, and 25%-83% of web applications. Active dynamic approaches have reported at least one misuse of security APIs, including crypto, SSL/TLS, OAuth, and Attestation APIs in 3%-41% of Android applications and 46%-58% of web applications.

Figure 11 illustrates different findings regarding the common misuse of security APIs. Insecure cryptography algorithms (broken hash) for crypto API, improper certificate validation for SSL API, misuse of the *State* parameter for OAuth API, and lack of cryptography for Fingerprint API are the misuses reported as the most frequent occurrences.

7.2 Performance Evaluation

Our review identified 25 studies that performed experiments to assess the performance of misuse detection techniques. Out of these, 5 studies [S31, S32, S34, S41, S42] evaluated the performance of various static analysis tools in detecting crypto and SSL/TLS misuses. Following we discuss various metrics, benchmarks, and strategies adopted for evaluation.

7.2.1 Evaluation Metrics. We have identified 10 evaluation metrics that are commonly used to measure the performance of security API misuse detection techniques. These metrics are grouped into two categories: *detection effectiveness* and *computation efficiency*.

Metrics for detection effectiveness are typically calculated using *True Positive (TP)*, *False Positive (FP)*, *True Negative (TN)*, or *False Negative (FN)* values. While detecting all misuses is crucial, having a high number of false alarms can be highly time-consuming and burdensome for developers. Hence, the primary objective of misuse detection is to maximize the *True Positive Rate (TPR)* or *Recall (R)*, while minimizing the *False Positive Rate (FPR)*. These two metrics are the most commonly used. Other metrics, such as *True Negative Rate (TNR)*, *False Negative Rate (FNR)*, *Precision (P)*, *Classification Accuracy*, *F-Measure*, and *Coverage* have also been used by researchers to measure the effectiveness of misuse detection.

Several primary studies also considered the computation efficiency of detection techniques, which was measured using the *runtime* and *space* complexity required for misuse detection. Evaluating computation efficiency is crucial in demonstrating the suitability of these techniques for real-world applications. Classification accuracy was used only by ML-based detection techniques, and coverage measurement is exclusive to dynamic analysis as it measures the proportion of the program code that has been executed during testing. Figure 12 illustrates the distribution of the identified evaluation metrics that are categorized by detection technique.

7.2.2 Evaluation Benchmarks. Benchmarks are critical for evaluating detection techniques and identifying their strengths and weaknesses. Unfortunately, there is an inadequate number of publicly available benchmarks, and all of them are limited to test cases for some misuses of Java crypto and SSL/TLS APIs. Table 7 lists 9 public benchmarks commonly used by researchers. Among them, the first five benchmarks were specifically created to evaluate and compare the performance of crypto misuse detection approaches.

CryptoAPI-Bench includes synthetic source codes with crypto API misuses, false positive tests, and correct API uses. It offers both basic test cases and advanced test cases that involve more complex scenarios. *CryptoAPI-Bench* was designed to assess static tools. *CryptoAPI-Bench** [S10] is an extension of *CryptoAPI-Bench* with additional cases suitable for evaluating dynamic approaches. *CryptoAPI-Bench* is not suitable for evaluating the scalability property of a tool, as all test cases are lightweight by design. To address this limitation, another study [S32] created *ApacheCryptoAPI-Bench* using 10 real-world Apache projects that are complex programs with numerous and lengthy code files. This benchmark is therefore appropriate for assessing the scalability and applicability of existing approaches to real-world applications.

Two additional benchmarks for evaluating crypto misuse detection techniques are *Braga et al.'s* [68] and *Fischer et al.'s* [67] datasets that contain labeled instances of secure and insecure use of the Java Cryptography Architecture API. Braga et al.'s dataset consists of synthetic Java source codes, while the latter consists of real-world code snippets collected from SO. Both datasets were used by the study [S65] to train and test an ML-based detection technique. The last four benchmarks have been designed to evaluate API misuse or vulnerability detectors and include some test cases for evaluating crypto misuse detection techniques as well.

7.2.3 Evaluation Strategies. In our review, various techniques were used to evaluate the performance of misuse detection models. We noticed only seven studies designed experiments using *public benchmarks*, as was shown in Table 7. Existing benchmarks are typically limited in terms of scale and diversity of test cases. To address this issue, one study [S34] explored the *automatic generation of test cases* using mutation operations. It generated over 20,000 test cases, which were used to evaluate several crypto misuse detection tools and identify their flaws, such as failure to detect insecure algorithms provided in lowercase. In addition, 19 primary studies conducted *case studies* or *manually analyzed* a subset of their dataset or a subset of reported misuses to verify their results. For instance, one study [S10] randomly selected 150 Android apps out of 1,780 analyzed apps to validate their findings, and another study [S58] randomly sampled 157 misuses and manually verified them to gain a deeper understanding of common false positives.

Table 7. Public benchmarks for evaluating crypto misuse detection techniques used by the reviewed studies

Benchmark	Size	Type	Description	Ref
CryptoAPI-Bench (2019) [S1][69]	181 test cases	Synthetic	Benchmark for evaluating crypto misuse detectors containing 45 basic and 136 complex test cases with crypto API misuses, false positive tests, and correct API uses	S1, S2, S31, S32
CryptoAPI-Bench* (2021) [S10]	198 test cases	Synthetic	CryptoAPI-Bench with further cases suitable for assessing dynamic approaches, totally consisting of 157 crypto misuse cases, and 41 normal test cases	S10
ApacheCryptoAPI-Bench (2020) [S32][70]	120 test cases	Real	Ten real-world Apache including 79 basic test cases and 42 advanced test cases, suitable for assessing the scalability of misuse detection approaches	S32
Braga et al.'s dataset (2017) [68]	384 test cases	Synthetic	Contains 202 misuses (positive cases) and 182 normal uses (negative cases) for Java Cryptography Architecture	S41, S65
Fischer et al.'s dataset (2019) [67]	16,346 test cases	Real	6,246 secure cases and 10,100 insecure cases for the use of crypto API adopted from code snippets available SO posts	S65
MUBench (2016) [71][72]	21 apps	Real	Benchmark for evaluating API-misuse detectors containing instances of crypto API misuses collected from 62 Java programs	S31
OWASP (2021) [73]	975 programs	Real	Java test suite designed for evaluating vulnerability detectors, containing 477 programs with labeled misuses of security APIs and 498 programs with correct uses	S31
DroidBench (2015) [74]	21 apps	Real	Benchmark apps for evaluating the performance of static information-flow analysis of Android apps including crypto misuse test cases	S44
ICC-Bench (2017) [75]	24 apps	Real	Benchmark apps for evaluating the performance of static analysis to detect inter-component data leakage problem of Android apps including crypto misuse test cases	S44

The study [S45] created a dataset for 5-fold cross-validation by manually labeling a collection of security-related code snippets from SO as either secure or insecure that were used for evaluating its proposed ML-based algorithms. Unlike the study [S45], the study [S18] relied on an *existing tool*, CogniCrypt_{SAST} [S11], to label crypto API use cases in Android applications and provide a labeled dataset for training, validating, and testing purposes of its ML algorithms, which makes the results dependent on the performance of the employed tool.

Another evaluation technique, adopted by 3 studies, involves *executing attacks* to validate the results and identify exploitable misuses. For instance, one study [S6] executed two ethical attacks on two applications and successfully retrieved personal information encrypted and transmitted over the network. Meanwhile, 16 studies disclosed misuses identified in real-world projects, some of which analyzed the feedback they received from developers. This analysis provided valuable insights into developers' requirements from misuse detection tools disregarded in existing approaches.

Lastly, we found 3 studies that conducted *user studies* to evaluate the usability of tools with warning messages and suggestions for fixes. These studies involved 39 developers [S5], 8 developers [S42], and 53 developers [S59] and showed that security advice could improve the usage of crypto APIs in users' codes. More importantly, they highlighted the need for detailed and specific solutions that are comprehensible and feasible for developers.

8 OPEN RESEARCH CHALLENGES AND FUTURE DIRECTIONS

This section provides recommendations for best practices and highlights areas that require further research based on the key findings of our SLR.

8.1 Need for research on non-cryptography APIs

While the focus of primary studies on cryptography and SSL APIs highlights their crucial role in secure software development, it raises a warning alarm about potential vulnerabilities caused by misusing the security APIs that receive inadequate or even no attention from the research community. Meng et al.'s analysis of SO posts [14] revealed that Spring Security is the most popular option among Java developers in secure coding practices. However, our review found only one study dedicated to this overly complicated and poorly documented API. On the other hand, research suggests that misusing other security APIs, including those designed for storing and accessing sensitive information, can lead to significant security implications [76–78]. As a result, developing effective mechanisms for detecting and preventing API misuse in these contexts is critical. Moreover, novel security APIs have been developed to facilitate the

integration of applications with cutting-edge security technologies, including various ML-based authentication schemes such as facial recognition, vocal recognition, and iris-based authentication, which are increasingly utilized in mobile applications [79]. However, there is a lack of research on the current state of ML-based security APIs, particularly concerning their usability and usage patterns. Therefore, conducting timely studies in these areas is imperative to prevent any serious consequences arising from their misuse before they are widely adopted.

8.2 Need for support of diverse programming languages

The complexity of API designs in Java and its popularity, especially in Android app development, motivated many researchers to examine the usage patterns of security APIs in Java-based applications. However, it is essential to recognize the significance of exploring this area and addressing the existing challenges for APIs in other widely used programming languages such as Python and C. Several primary studies conducted on non-Java languages indicate that developers in other languages are also likely to make mistakes when using security APIs. In addition to the need for developing tools for various languages, there is a special need for developing cross-language tools. For example, a study [S20] analyzing MicroPython projects suggests that developers working in the embedded domain often use C language to implement cryptography operations, highlighting the need for cross-language analysis capability for tracking program information across multiple programming languages. Furthermore, Meng et al.'s analysis of SO posts [14] has shed light on the challenge of cross-language data handling of cryptography APIs, where developers struggle to encrypt data in one language (e.g., PHP or Python) and decrypt data in another language (e.g., Java). Developing cross-language approaches can ensure more secure software development across different programming languages. Therefore, it is highly recommended that researchers broaden their research scope to include non-Java languages and focus on developing cross-language tools and approaches for security APIs.

8.3 Need for research on human-centric models

There is a significant research gap when it comes to understanding the human needs in misuse detection models for security APIs. Sixteen studies disclosed misuses identified in real-world projects, some of which [e.g., S1, S10, S31] reported the feedback received from developers. These studies revealed inadequacies of current tools in meeting the expectations and requirements of developers, indicating the need for adapting these tools based on user feedback and preferences. In certain cases, developers acknowledged the existence of misuses (71% [S21], 52% [S9], 20% [S29] of reported misuses) and attempted to address them. However, in some cases, they disregarded the identified misuses, believing them to be irrelevant. These misuses were often found in non-sensitive contexts such as security-irrelevant pieces of code, test cases, or archived code [S10, S31]. Additionally, in some cases, developers rejected misuses without concrete exploit demonstrations [S1, S10, S31]. To better align with developers' requirements, detection tools should be capable of differentiating between security-relevant and security-irrelevant contexts and demonstrating concrete exploit examples for identified misuses.

To fill this research gap, it is recommended that researchers conduct studies to understand and identify user requirements and expectations from misuse detection models in the context of security APIs. These studies should focus on gathering insights directly from developers, enabling a deeper understanding of their needs. Additionally, tool developers should prioritize user needs by adopting human-centric policies and incorporating user feedback into the design and development process. By addressing the human-centric aspects of misuse detection models, researchers, and tool developers can enhance the usability, effectiveness, and adoption of these tools among developers.

On the other hand, effective misuse detection is the first step toward the subsequent mitigation strategies for supporting developers in the effective use of security APIs. Developers may face challenges while trying to fix identified issues, which can result in the introduction of new mistakes [S42]. Thus, misuse detection tools need to be complemented by comprehensible, detailed, and actionable fixing suggestions. Based on the vulnerability disclosures reported in the reviewed studies, there were instances where developers acknowledged the misuses but were unable to resolve them. Some developers cited operational limitations, such as the need to maintain backward compatibility for clients, as hindering them from making necessary fixes [S1]. Others noted that the guidance offered by the tools was inadequate and failed to provide all the necessary details for repairs [S31]. Additionally, some developers found it challenging to deal with the complexity of implementing secure solutions [S31]. In our review, a few studies [S5, S59, S63] provide users with fix suggestions and general guidance for some crypto misuses, but they often lack customizing fixes for a given vulnerable program. There are only four studies [S2, S16, S17, S46] proposed automated generation of customized fixes for crypto misuses. Existing tools are still inadequate in assisting developers with accurately correcting misuses, highlighting the crucial need for more detailed and customized suggestions for repair [S31].

8.4 Need for applying state-of-the-art ML-based techniques

One of the key findings of our SLR is that there is a scarcity of state-of-the-art ML-based and extensible detection models for identifying security API misuses. Most primary studies rely on traditional heuristic-based approaches that require significant domain knowledge and result in labor-intensive, time-consuming, and error-prone processes. Manual analysis of identified misuses in one study [S58], using CrySL [S11], also revealed several false positives due to incorrect specifications in CrySL. Additionally, these methods often rely on hard-coded rules, making them difficult to adapt to emerging misuses to keep pace with the rapidly evolving security threat landscape.

Despite the availability of vast open-source repositories, which has motivated many researchers to explore data-driven techniques and ML and Deep Learning (DL) models to discover vulnerability patterns [80], there are few studies that have employed ML to implement misuse detection for security APIs. Furthermore, while there has been increasing interest in utilizing NLP techniques to understand source code semantics and patterns, the application of NLP in security API misuse detection remains largely unexplored. In addition, recent advancements in Large Language Models (LLMs), like ChatGPT, have shown remarkable performance in various tasks and attracted considerable attention from researchers in different domains. LLMs have demonstrated their capability to understand and generate both natural language and code, making them potential tools for addressing security API misuses. However, their specific application in this context is yet to be explored.

The integration of NLP, ML, and DL techniques can offer great promise in improving the accuracy and efficiency of security API misuse detection. Further research in this area, particularly providing required labeled data, could lead to more effective and adaptable solutions for identifying and mitigating security API misuses in the future. Additionally, the continuous adaptation of ML-based models with new data is essential to maintain their effectiveness and relevance in the face of evolving security APIs, misuses, and trends.

8.5 Need for developing evaluation benchmarks

Benchmarks are crucial to compare and evaluate the performance of proposed approaches and tools for security API misuse detection and to identify the areas for further improvement. However, our review identified only 8 publicly available benchmarks, out of which only five were specially designed for security API misuses. All these benchmarks are limited to test cases for some misuses of crypto or SSL/TLS APIs in the Java programming language. Thus, there is

a desperate need to expand the scope of benchmarks to cover a broader range of security APIs, real-world misuses, programming languages, and software platforms suitable for evaluating tools developed for diverse security APIs in real-world scenarios. Although some studies have manually analyzed a random subset of misuses identified in their experiments, none of them have publicly shared the results of their analysis. Therefore, we recommend that datasets resulting from manual analysis be made available to support validation and future research. On the other hand, existing benchmarks are mostly limited to test cases for evaluating static tools, which renders them unsuitable for assessing the performance of dynamic analysis tools [S10]. Therefore, benchmarks need to come with test cases for evaluating dynamic analysis tools. It is also crucial to continuously update existing benchmarks to incorporate new misuses and the future evolution of APIs. Furthermore, we recommend versioning benchmarks to address issues such as concept and temporal drift.

9 THREATS TO VALIDITY

We followed the guidelines outlined in study [26] to design and conduct our SLR. We took necessary steps to minimize the impact of any potential threats to the validity of the SLR, which are elaborated upon below:

One of the common threats to the validity of an SLR is the possibility of missing relevant studies. To minimize this risk, we utilized Scopus, which is the most comprehensive search engine and largest indexing system [27, 28], and supplemented it with the two most frequently used digital libraries, IEEE Xplore and ACM Digital Library [29]. We also conducted a series of pilot searches to establish a search string that would retrieve relevant papers already known to us. In addition, we employed both forward and backward snowballing techniques to locate any other relevant papers that might have been missed by the search string.

The potential for subjective bias in the selection of studies cannot be ruled out, as it could be influenced by the author's subjective judgment. To address this concern, we carried out a rigorous and well-defined multi-step process (detailed in Section 3.3) with clear inclusion and exclusion criteria. We also established specific guidelines to exclude low-quality papers. At every stage of the selection process, we carefully deliberated and addressed any uncertainties to minimize the risk of selection bias.

Human errors and author biases during data extraction, analysis, and interpretation can impact the accuracy of the results and findings. To mitigate this issue, a data extraction form was developed and refined to ensure the collection of adequate and consistent information for answering the research questions. We also conducted fortnightly meetings to review and verify the synthesis and interpretation of our quantitative and qualitative analysis and to resolve any disagreements before finalizing our responses to the research questions.

10 CONCLUSION

Security APIs play a crucial role in secure software development. Prior studies have shown developers often misuse security APIs, leading to costly software vulnerabilities. Thus, misuse detection for security APIs has gained significant attention from the research community for ensuring software security. However, the existing literature on the topic is dispersed, and a systematic review was necessary to identify the state-of-the-art approaches and highlight areas that require further exploration. This study presents our research effort aimed at systematically reviewing and rigorously analyzing the literature on misuse detection for security APIs. To the best of our knowledge, this SLR is the first attempt to systematically review the literature on this topic. We have provided an organized evidence-based body of knowledge to enrich this domain by identifying security APIs, their potential misuses, detection techniques, and evaluation methods. In conclusion, based on a comprehensive analysis of 69 primary studies, we identified key trends in security API misuse detection research that are:

- 1) We identified 6 security APIs examined for misuse detection, namely cryptography primitives (crypto), SSL/TLS, OAuth, Fingerprint, Spring, and SafetyNet Attestation. Most studies focused on crypto and SSL/TLS, highlighting the need to explore this topic for other security APIs.
- 2) We identified a total of 39 misuses, including 6 crypto, 6 SSL/TLS, 10 OAuth, 4 Fingerprint, 4 Spring, and 9 SafetyNet Attestation misuses. The primary studies mainly focused on analyzing Android apps, and the most commonly reported misuses were using insecure crypto algorithms for crypto APIs, improper certificate validation for SSL APIs, misuse of the *state* parameter for OAuth APIs, and lack of cryptography for Fingerprint APIs.
- 3) We proposed a taxonomy consisting of heuristic-based and ML-based approaches for misuse detection techniques. Most studies relied on heuristic-based approaches, with 42 studies based on static analysis, 9 studies using dynamic analysis, and 13 studies using a hybrid approach. We found only 3 studies using ML to address misuse detection. Our findings suggest the need to explore the application of ML, DL, and NLP techniques in this area.
- 4) We identified 11 metrics for evaluation, grouped into accuracy and efficiency categories. We found only five public benchmarks, particularly designed for security API misuse, which are limited to test cases for crypto and SSL/TLS misuses. These findings highlight the need for further research and development of more diverse benchmarks to facilitate the evaluation of misuse detection techniques for security APIs.

Overall, our review offers valuable insights for both researchers and practitioners. Researchers can leverage the research gaps, taxonomies of misuses and detection techniques to advance their research. Particularly, our research highlighted a crucial area that needs attention - understanding what developers expect and require from misuse detection models. Filling this gap results in significant advancement in this field. Practitioners can also benefit from our findings by selecting appropriate techniques, improving their tools through best practices, and adopting human-centric policies.

ACKNOWLEDGMENTS

The work has been supported by the Cyber Security Research Centre Limited whose activities are partially funded by the Australian Government's Cooperative Research Centres Programme.

REFERENCES

- [1] IBM. Cost of a data breach 2022 report, 2022. URL <https://www.ibm.com/reports/data-breach>. Accessed June 10, 2023.
- [2] Herb Krasner. The cost of poor software quality in the us: A 2020 report. *Proc. Consortium Inf. Softw. QualityTM (CISQTM)*, 2021.
- [3] Equifax. Data breach notice, 2017. URL <https://www.equifaxsecurity2017.com/consumer-notice/>. Accessed June 13, 2023.
- [4] Optus. Data breach notice, 2022. URL <https://www.optus.com.au/about/media-centre/media-releases/2022/09/optus-notifies-customers-of-cyberattack>. Accessed June 13, 2023.
- [5] Chamila Wijayarathna and Nalin Asanka Gamagedara Arachchilage. Using cognitive dimensions to evaluate the usability of security APIs: an empirical investigation. *Information and Software Technology*, 115:5–19, 2019.
- [6] Stefan Krüger, Johannes Späth, Karim Ali, Eric Bodden, and Mira Mezini. CrySL: An extensible approach to validating the correct usage of cryptographic APIs. *IEEE Transactions on Software Engineering*, 47(11):2382–2400, 2019.
- [7] Martin Georgiev, Subodh Iyengar, Suman Jana, Rishita Anubhai, Dan Boneh, and Vitaly Shmatikov. The most dangerous code in the world: validating SSL certificates in non-browser software. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 38–49, 2012.
- [8] David Sounthiraraj Justin Sahs Garret Greenwood and Zhiqiang Lin Latifur Khan. SMV-hunter: Large scale, automated detection of SSL/TLS Man-in-the-Middle vulnerabilities in Android apps. In *Network and Distributed System Security Symposium (NDSS)*, pages 1–14, 2014.
- [9] Antonio Bianchi, Yanick Fratantonio, Aravind Machiry, Christopher Kruegel, Giovanni Vigna, Simon Pak Ho Chung, and Wenke Lee. Broken Fingers: On the Usage of the Fingerprint API in Android. In *NDSS*, 2018.
- [10] Mohammadreza Hazhirpasand, Mohammad Ghafari, Stefan Krüger, Eric Bodden, and Oscar Nierstrasz. The impact of developer experience in using Java cryptography. In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–6. IEEE, 2019.
- [11] Berk Sunar. True random number generators for cryptography. *Cryptographic Engineering*, pages 55–73, 2009.
- [12] Matthew Green and Matthew Smith. Developers are not the enemy!: The need for usable security APIs. *IEEE Security & Privacy*, 14(5):40–46, 2016.
- [13] Yasemin Acar, Christian Stransky, Dominik Wermke, Charles Weir, Michelle L Mazurek, and Sascha Fahl. Developers need support, too: A survey of security advice for software developers. In *2017 IEEE Cybersecurity Development (SecDev)*, pages 22–26. IEEE, 2017.

- [14] Na Meng, Stefan Nagy, Danfeng Yao, Wenjie Zhuang, and Gustavo Arango Argoty. Secure coding practices in Java: Challenges and vulnerabilities. In *Proceedings of the 40th International Conference on Software Engineering*, pages 372–383, 2018.
- [15] Roland Croft, Yongzheng Xie, Mansoor Zahedi, M Ali Babar, and Christoph Treude. An empirical study of developers' discussions about security challenges of different programming languages. *Empirical Software Engineering*, 27:1–52, 2022.
- [16] Triet Huynh Minh Le, Roland Croft, David Hin, and Muhammad Ali Babar. A large-scale study of security vulnerability support on developer q&a websites. In *Evaluation and assessment in software engineering*, pages 109–118. 2021.
- [17] Felix Fischer, Konstantin Böttinger, Huang Xiao, Christian Stransky, Yasemin Acar, Michael Backes, and Sascha Fahl. Stack Overflow considered harmful? the impact of copy&paste on Android application security. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 121–136. IEEE, 2017.
- [18] Ying Zhang, Md Mahir Asef Kabir, Ya Xiao, Danfeng Yao, and Na Meng. Automatic detection of Java cryptographic API misuses: Are we there yet? *IEEE Transactions on Software Engineering*, 49(1):288–303, 2022.
- [19] Maxime Lamothe, Yann-Gaël Guéhéneuc, and Weiyi Shang. A systematic review of API evolution literature. *ACM Computing Surveys (CSUR)*, 54(8):1–36, 2021.
- [20] Peter Leo Gorski and Luigi Lo Iacono. Towards the usability evaluation of security APIs. In *Clarke, Furnell (Eds.): Tenth International Symposium on Human Aspects of Information Security & Assurance (HAISA 2016), Frankfurt, Germany, July 19-21, 2016*, pages 252–265. CSCAN, 2016.
- [21] Michael E Whitman and Herbert J Mattord. *Management of information security*. Cengage Learning, 2013.
- [22] Dick Hardt. RFC 6749: The OAuth 2.0 authorization framework, 2012.
- [23] Ethan Shernan, Henry Carter, Dave Tian, Patrick Traynor, and Kevin Butler. More guidelines than rules: CSRF vulnerabilities from noncompliant OAuth 2.0 implementations. In *Detection of Intrusions and Malware, and Vulnerability Assessment: 12th International Conference, DIMVA 2015, Milan, Italy, July 9-10, 2015, Proceedings 12*, pages 239–260. Springer, 2015.
- [24] Hala Assal and Sonia Chiasson. Security in the software development lifecycle. In *SOUPS@USENIX Security Symposium*, pages 281–296, 2018.
- [25] Barbara A Kitchenham, Tore Dyba, and Magne Jorgensen. Evidence-based software engineering. In *Proceedings. 26th International Conference on Software Engineering*, pages 273–281. IEEE, 2004.
- [26] B. Kitchenham and S. Charters. Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, Ver. 2.3 EBSE Technical Report. EBSE, 2007.
- [27] Triet HM Le, Huaming Chen, and M Ali Babar. A survey on data-driven software vulnerability assessment and prioritization. *ACM Computing Surveys (CSUR)*, 2021.
- [28] Roland Croft, Yongzheng Xie, and Muhammad Ali Babar. Data preparation for software vulnerability prediction: A systematic literature review. *IEEE Transactions on Software Engineering*, 2022.
- [29] Guanjun Lin, Sheng Wen, Qing-Long Han, Jun Zhang, and Yang Xiang. Software vulnerability detection using deep neural networks: A survey. *Proceedings of the IEEE*, 108(10):1825–1848, 2020.
- [30] Bushra Sabir, Faheem Ullah, M Ali Babar, and Raj Gaire. Machine learning for detecting data exfiltration: a review. *ACM Computing Surveys (CSUR)*, 54(3):1–47, 2021.
- [31] Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, pages 1–10, 2014.
- [32] Z. Mousavi, C. Islam, M. A. Babar, A. Abuadbbba, and K. Moore. Online Appendix of "Detecting Misuses of Security APIs: A Systematic Review", 2023. URL <https://github.com/szmousavi/SLR-of-Security-API-Misuse-Detection>. Accessed June 10, 2023.
- [33] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.
- [34] OpenSSL. URL <https://www.openssl.org/>. Accessed June 10, 2023.
- [35] Nat Sakimura, John Bradley, Mike Jones, Breno de Medeiros, and Chuck Mortimore. OpenID Connect core 1.0 incorporating errata set 1, 2014.
- [36] YubiCo. Yubikeys, 2017. URL <https://www.yubico.com/products/yubikey-hardware/>. Accessed June 10, 2023.
- [37] Google. New in Android samples: Authenticating to remote servers using the fingerprint API, 2015. URL <https://android-developers.googleblog.com/2015/10/new-in-android-samples-authenticating.html>. Accessed June 10, 2023.
- [38] OWASP-MASTG. Local authentication on Android, 2017. URL <https://github.com/OWASP/owasp-mastg/blob/master/Document/0x05f-Testing-Local-Authentication.md>. Accessed June 10, 2023.
- [39] Spring. Spring security. URL <https://spring.io/projects/spring-security>. Accessed June 10, 2023.
- [40] Google. SafetyNet Attestation API, 2020. URL <https://developer.android.com/training/safetynet/attestation>. Accessed June 10, 2023.
- [41] Elaine Barker, William Burr, Alicia Jones, Timothy Polk, Scott Rose, Miles Smid, Quynh Dang, et al. Recommendation for key management part 3: Application-specific key management guidance. *NIST special publication*, 800:57, 2009.
- [42] Oracle. Jdk 19 documentation, 2022. URL <https://docs.oracle.com/en/java/javase/19/>. Accessed June 10, 2023.
- [43] Elaine Barker and Quynh Dang. Nist special publication 800-57 part 1, revision 4. *NIST, Tech. Rep.*, 16, 2016.
- [44] Daniel J Bernstein, Yun-An Chang, Chen-Mou Cheng, Li-Ping Chou, Nadia Heninger, Tanja Lange, and Nicko Van Someren. Factoring rsa keys from certified smart cards: Coppersmith in the wild. In *Advances in Cryptology-ASIACRYPT 2013: 19th International Conference on the Theory and Application of Cryptology and Information Security, Bengaluru, India, December 1-5, 2013, Proceedings, Part II 19*, pages 341–360. Springer, 2013.
- [45] Hugo Krawczyk. How to predict congruential generators. *Journal of algorithms*, 13(4):527–545, 1992.
- [46] Burt Kaliski and A Rusch. RFC 8018: PKCS# 5: Password-based cryptography specification version 2.1, 2017.

- [47] Richard Barnes, Martin Thomson, Alfredo Pironti, and Adam Langley. Deprecating secure sockets layer version 3.0, 2015. URL <https://tools.ietf.org/html/rfc7568>.
- [48] K. Moriarty and S. Farrell. Deprecating TLSv1.0 and TLSv1., 2021. URL <https://tools.ietf.org/html/draft-ietf-tls-oldversions-deprecate-12>. Accessed June 10, 2023.
- [49] Sean Turner and Tim Polk. Prohibiting secure sockets layer (SSL) version 2.0. Technical report, 2011.
- [50] Moxie Marlinspike. More tricks for defeating SSL in practice. *Black Hat USA*, 516, 2009.
- [51] Moxie Marlinspike. New tricks for defeating SSL in practice. *Black Hat DC*, 2, 2009.
- [52] Tamjid Al Rahat, Yu Feng, and Yuan Tian. Cerberus: Query-driven scalable vulnerability detection in OAuth service provider implementations. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2459–2473, 2022.
- [53] Nat Sakimura, John Bradley, and Naveen Agarwal. Proof key for code exchange by OAuth public clients. Technical report, 2015.
- [54] Common Weakness Enumeration. CWE-306: Missing Authentication for Critical Function, . URL <https://cwe.mitre.org/data/definitions/306.html>. Accessed June 10, 2023.
- [55] Common Weakness Enumeration. CWE-862: Missing authorization, . URL <https://cwe.mitre.org/data/definitions/862.html>. Accessed June 10, 2023.
- [56] Common Weakness Enumeration. CWE-863: Incorrect authorization, . URL <https://cwe.mitre.org/data/definitions/863.html>. Accessed June 10, 2023.
- [57] Marc Stevens, Elie Bursztein, Pierre Karpman, and Ange Albertini. Yarik markov, alex petit bianco, and clement baisse. announcing the first sha1 collision. *Google Security Blog*, <https://security.googleblog.com/2017/02/announcing-first-sha1-collision.html>, 2017.
- [58] Patrick Lam, Eric Bodden, Ondrej Lhoták, and Laurie Hendren. The Soot framework for Java program analysis: a retrospective. In *Cetus Users and Compiler Infrastructure Workshop (CETUS 2011)*, volume 15, 2011.
- [59] Tumbleson Connor and Wiśniewski Ryszard. Apktool - a tool for reverse engineering Android apk files, 2010. URL <https://ibotpeaches.github.io/Apktool/>. Accessed June 10, 2023.
- [60] Anthony Desnos. Reverse engineering and pentesting for Android applications, 2012. URL <https://github.com/androguard/androguard>. Accessed June 10, 2023.
- [61] IBM: TJ. Watson libraries for analysis WALA. URL <https://wala.sourceforge.net/>. Accessed June 10, 2023.
- [62] Harshal Tupsamudre, Monika Sahu, Kumar Vidhani, and Sachin Lodha. Fixing the fixes: Assessing the solutions of sast tools for securing password storage. In *Financial Cryptography and Data Security: FC 2020 International Workshops, AsiaUSEC, CoDeFi, VOTING, and WTSC, Kota Kinabalu, Malaysia, February 14, 2020, Revised Selected Papers 24*, pages 192–206. Springer, 2020.
- [63] Baowen Xu, Ju Qian, Xiaofang Zhang, Zhongqiang Wu, and Lin Chen. A brief survey of program slicing. *ACM SIGSOFT Software Engineering Notes*, 30(2):1–36, 2005.
- [64] Steven Arzt, Siegfried Rasthofer, Christian Fritz, Eric Bodden, Alexandre Bartel, Jacques Klein, Yves Le Traon, Damien Oceau, and Patrick McDaniel. Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for Android apps. *Acm Sigplan Notices*, 49(6):259–269, 2014.
- [65] James C King. Symbolic execution and program testing. *Communications of the ACM*, 19(7):385–394, 1976.
- [66] Kevin Allix, Tegawendé F Bissyandé, Jacques Klein, and Yves Le Traon. Androzo: Collecting millions of Android apps for the research community. In *Proceedings of the 13th international conference on mining software repositories*, pages 468–471, 2016.
- [67] Felix Fischer, Huang Xiao, Ching-Yu Kao, Yannick Stachelscheid, Benjamin Johnson, Danial Razar, Paul Fawkesley, Nat Buckley, Konstantin Böttinger, Paul Muntean, et al. Stack Overflow considered helpful! deep learning security nudges towards stronger cryptography. In *USENIX Security Symposium*, pages 339–356, 2019.
- [68] Alexandre Braga, Ricardo Dahab, Nuno Antunes, Nuno Laranjeiro, and Marco Vieira. Practical evaluation of static analysis tools for cryptography: Benchmarking method and case study. In *International Symposium on Software Reliability Engineering (ISSRE)*, pages 170–181. IEEE, 2017.
- [69] CryptoAPI-Bench, 2019. URL <https://github.com/CryptoGuardOSS/cryptoapi-bench>. Accessed June 10, 2023.
- [70] ApacheCryptoAPI-Bench, 2020. URL <https://github.com/CryptoAPI-Bench/ApacheCryptoAPI-Bench>. Accessed June 10, 2023.
- [71] Sven Amann, Sarah Nadi, Hoan A Nguyen, Tien N Nguyen, and Mira Mezini. MUBench: A benchmark for API-misuse detectors. In *Proceedings of the 13th international conference on mining software repositories*, pages 464–467, 2016.
- [72] MUBench, 2016. URL <https://GitHub.com/stg-tud/MUBench>. Accessed June 10, 2023.
- [73] OWASP Benchmark, 2021. URL <https://owasp.org/www-project-benchmark/>. Accessed June 10, 2023.
- [74] DroidBench, 2015. URL <https://GitHub.com/secure-software-engineering/DroidBench>. Accessed June 10, 2023.
- [75] ICC-Bench, 2017. URL <https://GitHub.com/fgwei/ICC-Bench>. Accessed June 10, 2023.
- [76] Amiangshu Bosu, Fang Liu, Danfeng Yao, and Gang Wang. Collusive data leak and more: Large-scale threat analysis of inter-app communications. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 71–85, 2017.
- [77] Yuhong Nan, Zheming Yang, Xiaofeng Wang, Yuan Zhang, Donglai Zhu, and Min Yang. Finding clues for your secrets: Semantics-driven. *Learning-Based Privacy Discovery in Mobile Apps*, 10, 2018.
- [78] Chaoshun Zuo, Zhiqiang Lin, and Yinqian Zhang. Why does your data leak? uncovering the data leakage in cloud from mobile apps. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1296–1310. IEEE, 2019.
- [79] Douglas Kunda and Mumbi Chishimba. A survey of Android mobile phone authentication schemes. *Mobile Networks and Applications*, 26(6): 2558–2566, 2021.
- [80] Saikat Chakraborty, Rahul Krishna, Yangruibo Ding, and Baishakhi Ray. Deep learning based vulnerability detection: Are we there yet? *IEEE Transactions on Software Engineering*, 2021.