Comprehensive Quality Assessment for Security Vulnerability Databases

ABSTRACT

Security vulnerability databases play a vital role in collecting, categorizing, and revealing vulnerabilities to the public. However, previous research has revealed significant inconsistencies across these databases and some have been found to contain inaccurate information or to be missing.

In our paper, we propose CQASVD, a comprehensive quality assessment system designed to evaluate the data quality of various vulnerability databases. CQASVD mainly consists of a comprehensive data quality model and a comprehensive assessment model. The comprehensive data quality model defines the dimensions, while the comprehensive assessment model employs various Multi-Criteria Decision-Making (MCDM) methods to assess the quality of the databases. As far as we know, CQASVD is the first tool for assessing the quality of vulnerability databases. Our evaluation of various MCDM methods indicates that NVD, SecurityFocus, and CNNVD are highly comprehensive databases, and NVD remains the top vulnerability database, with SecurityFocus having a significant impact on all vulnerability databases.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Vulnerability Databases, Data Quality, Quality Assessment

ACM Reference Format:
. 2018. Comprehensive Quality Assessment for Security Vulnerability Databases. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Different organizations have developed vulnerability databases to address the growing number of cyberattacks. For instance, the National Vulnerability Database (NVD)[4] collects information on over 150,000 security vulnerabilities, which significantly alleviates the harm caused by security vulnerabilities.

However, several recent studies have pointed out vulnerability databases' data inconsistencies and inaccurate information. For example, Chaparro et al.[10] have focused on identifying missing steps to reproduce vulnerabilities. Nguyen et al.[18], and Nappa et al.[17] focus on vulnerable versions in NVD, and Dong et al.[13] address inconsistencies in versions across different vulnerability databases. However, The following questions remain difficult to answer: *Given two vulnerability databases, which one has higher quality and is more authoritative?*

The paper presents CQASVD, a comprehensive quality assessment system specifically designed for evaluating vulnerability databases, as a solution to the aforementioned questions. Data quality defined in the ISO/IEC 25012 standard[14] does not fully consider the specific characteristics of vulnerability databases. CQASVD selected relevant data dimensions to establish a comprehensive quality model suitable for vulnerability databases. Additionally, CQASVD's comprehensive assessment model uses various Multi-Criteria Decision-Making (MCDM) methods to generate quality rankings for different vulnerability databases.

In sum, the main contributions of this paper are as follows:

- First, we design and implement a method called CQASVD to build a comprehensive quality model for vulnerability databases for the first time.
- Second, we designed and implemented a prototype of the comprehensive quality assessment system in CQASVD. We use CQASVD to evaluate the comprehensive quality of different vulnerability databases.
- The experimental results demonstrate that NVD, SecurityFocus, and CNNVD are highly comprehensive databases. Among these, NVD is the best option despite its several shortcomings. SecurityFocus has a significant impact on various vulnerability databases, while CNNVD is consistent with NVD.

2 BACKGROUND AND CHALLENGES

Security Vulnerability Databases. We select vulnerability databases for comprehensive quality assessment based on three criteria: (a) wide recognition and popularity, (b) a significant number of vulnerability entries (databases with less than 10,000 entries were disregarded), and (c) easy access to the database and structured data. The vulnerability databases chosen include NVD[4], Exploit DB (EDB)[3], SecurityFocus (SF)[5], CNNVD[2], SeeBug[7], SecurityTracker (ST)[6], Vigil@nce[8], and RNVD[1]. We also collected related files, which helped us understand how the vulnerability database handles vulnerabilities.

Data Quality. The lack of appropriate definitions for data quality in vulnerability databases is likely due to database's unique characteristics. According to ISO/IEC 25012[14], data quality is the degree to which data meets explicit and implicit requirements when used under specified conditions. However, in this paper, we redefined the relevant data dimensions to align with the characteristics of vulnerability databases.

Challenges. Our goal is to achieve a comprehensive data quality assessment of different vulnerability databases. However, this

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

presents several challenges: first, the distinct nature of vulnerability databases necessitates a specialized data quality model tailored to their characteristics. Second, selecting the right Multi-Criteria Decision-Making (MCDM) methods to perform the comprehensive quality evaluation is a challenge. Lastly, differences in field structures, processing methods, and trends among databases can greatly impact the assessment, requiring the normalization of information.

3 THE DESIGN OF CQASVD

Figure 1 illustrates the overall process of CQASVD, a comprehensive quality assessment system consisting of three parts: vulnerability data and files acquisition, comprehensive quality model, and comprehensive evaluation model.

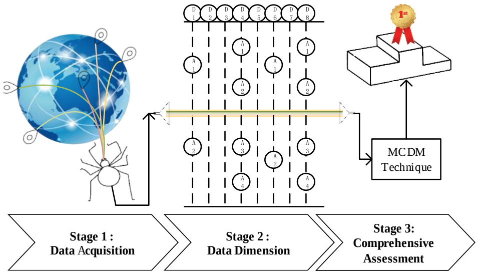

Figure 1: The Overall Process of CQASVD.

A. Comprehensive Quality Model. The specialized data quality model we developed for vulnerability databases is depicted in Figure 2. We selected four dimensions from the ISO/IEC 25012 standard[14]: Completeness, Accuracy, Compliance, and Availability. To attain a more comprehensive view of the performance of vulnerability databases, we added a new component, "Data Application and Analysis," which encompasses Data Volume, Timeliness, Influence, and Serviceability. We use a ratio score to evaluate the performance of various vulnerability databases in each data dimension.

B. Comprehensive evaluation model. Multi-Criteria Decision-Making (MCDM) focuses on organizing and resolving decision-making problems that involve multiple criteria. After creating the comprehensive quality model, we use the MCDM methods such as Analytic Hierarchy Process(AHP), Entropy-based, and Deviation Maximization to assess the vulnerability databases.

4 COMPREHENSIVE QUALITY ANALYSIS

The comprehensive quality model in Figure 2 is divided into two parts: **data quality** and **data analysis and application**. In this section, we redefine and analyze the related data dimensions. Table 1 presents the values for each data dimension for each dataset.

4.1 Data Quality

4.1.1 Completeness. We define two completeness: field completeness and information completeness.

Definition (Field Completeness). Fields are symbols that describe specific information about vulnerabilities, such as software names. Different databases use varying field types to describe related information. However, basic information about a vulnerability should be complete. This basic information is referred to as a fixed fieldset. We summarized various vulnerability databases and developed a fixed fieldset, which includes: **ID, CVE-ID, Title, Description, Entity, Operating System (OS), Vulnerability Type, Patch, Severity, Threat Type, Published Time, Verified, and Reference**. Field completeness assesses the extent to which the basic field types in a database match the fixed fieldset.

Analysis Our findings revealed inconsistencies in the field settings among different databases. For instance, NVD incorporates the Patch and Exploit fields into the Reference category. EDB mixes structured and unstructured vulnerability reports. As a commercial vulnerability database, Vigil@nce only discloses some simple information. Other vulnerability databases, such as CNNVD, SeeBug, RNVD, and ST, are already close to the fixed fieldset. A well-defined fieldset is necessary to enhance the efficiency of vulnerability information retrieval.

Definition (Information Completeness). Many databases are missing the corresponding content in specific fields, resulting in incomplete information. Information completeness refers to the extent to which all vulnerability entries contain complete information content, which reflects which information in the vulnerability library is easily overlooked.

Analysis Table 2 highlights the missing information in different vulnerability databases. The most commonly overlooked information is the CVE-ID, which may be due to the vulnerability database not having a complete mapping relationship with CVE or collecting vulnerabilities not recorded by the CVE. Patch information is also missing, as this is the developer's responsibility, not the vulnerability database, so patches receive less attention. Finally, the vulnerability type information is also frequently missing. Despite the Comprehensive Weakness Enumeration (CWE) providing a wide range of vulnerability types, it still lacks a thorough vulnerability type analysis. Information completeness reveals that missing data in vulnerability databases is a frequent occurrence and can range in degree.

4.1.2 Accuracy. Since security vulnerabilities are often tied to specific software, systems, and versions, it's important to describe them with clear and specific language to minimize ambiguity. Accuracy can then be divided into name accuracy and version accuracy.

Definition (Name Accuracy) Name accuracy measures the proportion of vulnerabilities that accurately describe the software name associated with the vulnerability. If the vulnerability information includes aliases, it can cause ambiguity. Abbreviations can also be considered a form of aliases. The likelihood of multiple applications sharing the same abbreviation increases, making it more challenging to handle vulnerabilities.

Analysis CNNVD, RNVD, SF, EDB, and Vigil@nce have high name accuracy, and these vulnerability databases all provide good standardized software names. However, the software name in SeeBug marks the affected components, and some even lack the software name. Only part of the ST database provides structured software names, and part is hidden in the title. CPE2.3, part of the SCAP protocol cluster, allows NVD to use abbreviations rather than official software names, resulting in a lower value for NVD.

Definition (Version Accuracy) The representation of software version information using terms such as $before$, $earlier$, and $-x$ is

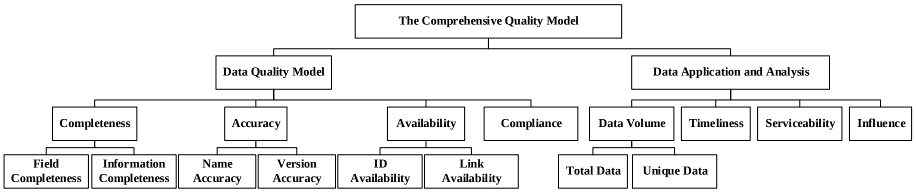

Figure 2: Comprehensive Quality Model of CQASVD.

Table 1: The different dimension values for each vulnerability databases.

Dims	Sub-Dims	NVD	CNNVD	RNVD	SF	ST	EDB	SeeBug	Vigil@nce
Completeness	Field-Com.	0.6923	0.8462	0.9231	0.7692	0.9231	0.6154	0.8462	0.6923
	Info-Com.	0.6808	0.4879	0.3564	0.317	0.194	0.486	0.0161	0.7931
Accuracy	Name-Acc.	0.482	0.82	0.963	0.979	0.53	0.969	0.208	0.824
	Version-Acc.	0.491	0.728	0.823	0.896	0.644	0.723	0.163	0.071
Availability	ID-Ava.	0.8099	0.9484	0.9999	0.9289	0.6254	0.9354	0.5737	0.9451
	Link-Ava.	2.8335	2.5832	1.9367	1.2259	2.0287	0	0.5708	2.876
Compliance	-	1	1	1	0.6666	0.3333	0.4444	1	0.4444
DataVolume	Total.	155,702	153,041	30,320	101,607	25,792	46,144	56,820	31,121
	Unique.	0	8,249	734	29,151	7771	20,220	38,614	6,435
Influence	-	0.8766	0.001	0.00001	1	0.2488	0.5379	0.0005	0.0516
Timeliness	-	0	-32.5869	-18.0675	-60.6275	-11.9232	-6.5431	2021.18	-24.8218
Serviceability	-	0.5	0.9	0.7	0.1	0	0.1	0.1	0.2

Table 2: Measurement results of information missing severity in vulnerability databases.

DBs	Top1(%)	Top2(%)	Top3(%)	Top4(%)	Top5(%)
CNNVD	Patch	Vul_type	Entity	CVE ID	Severity
RNVD	OS	Vul_type	Published	CVE ID	Verified
NVD	Vul_type	Ref	Entity	Severity_V2	
EDB	CVE ID	Verified			
SF	Patch	CVE ID	Vul_type	Entity	Ref
ST	CVE ID	OS	Patch	Published	Verified
SeeBug	Description	CVE ID	Entity	Ref	Patch
Vigil@nce	CVE ID	Modified	Ref		

inaccurate. This lack of precision makes it uncertain which versions may be affected by a vulnerability.

Analysis The SeeBug and Vigil@nce vulnerability databases rarely mention a vulnerability version, making it difficult to measure vulnerability version accuracy. NVD contains many words, such as $before$, and $earlier$, which leads to a large risk of version accuracy. The SF database lists the software version associated with each vulnerability and is known for its accuracy. CNNVD, RNVD, EDB, and others provide limited version information, so their version accuracy is better.

4.1.3 *Availability.*

The availability is classified into two categories: ID availability and link availability.

Definition (ID Availability). ID Availability measures the proportion of available vulnerabilities in a vulnerability database over a specified period. For instance, in CVE, a vulnerability that has been assigned an ID retains that ID even if verification fails, which results in an invalid vulnerability ID. ID Availability indicates the proportion of real vulnerabilities.

Analysis Except for ST and SeeBug, the data of other databases are relatively complete. CVE mainly calculates the ID of the NVD vulnerability database, and some CVE IDs may be revoked, reducing NVD's ID availability. However, ST and SeeBug databases are due to the loss of early data, resulting in unavailability.

Definition (Link Availability). However, many links in the invulnerability databases may become invalid over time. Hence, the number of valid reference links serves as a useful indicator.

Analysis The EDB database does not provide relevant links, so its link availability is 0. NVD, CNNVD, RNVD, ST, and Vigil@nce provide many links, and their average number is relatively high. SF is the first to publish information, so the number of links is limited as shown Timeliness in Table 1. SeeBug database rarely provides link information, which leads to low link availability.

4.1.4 *Compliance.*

Definition (Compliance). Compliance refers to the degree to which data conforms to established standards, conventions, or regulations in a specific use context. We have investigated and summarized the general processes for dealing with vulnerabilities in various databases and created a set of standards for vulnerability processing, which are illustrated in Table 3.

Our evaluation covered sub-dimensions such as **Submission, Secure transmission, Code, Audition&Processing, Grading, Name, Classification, Description, Entity**. We determined the support of these sub-dimensions by consulting publicly available resources and documents. We determine compliance by comparing the specifications of the vulnerability database under test with the specifications we have defined.

Analysis Vulnerability databases built by government agencies will always be more standardized than databases built by other agencies in Table 3. Different vulnerability databases have varying standardization requirements due to their differing focuses. EDB prioritizes PoCs and Vigil@nce prioritizes Patches, while SeeBug and SF aim for comprehensive vulnerability databases, but with limited success. ST faces numerous issues and fails to meet necessary normalization requirements. Each vulnerability database has its own unique standards and methods for expressing information. For example, NVD uses CWE to categorize vulnerability types and relies on CVSS to measure vulnerability severity, while SF and ST have their own expression systems and CNNVD and Vigil@nce have their own grading methods.

Table 3: Compliance results of the different vulnerability databases.

DBs	CNNVD	RNVD	NVD	EDB	SF	ST	SeeBug	Vigil@nce
Submission	○	○	○	○	○	×	○	○
Secure transmission	○	○	○	×	×	×	○	×
Code	○	○	○	○	○	○	○	○
Audition &Processing	○	○	○	○	○	×	○	○
Name	○	○	○	○	○	×	○	○
Grading	○	○	○	×	×	○	○	×
Description	○	○	○	×	×	×	○	×
classification	○	○	○	×	○	○	○	×
Entity	○	○	○	×	○	×	○	×

4.2 Data Analysis and Application

To accurately assess the overall performance of vulnerability databases, we have defined a component called "Data Application and Analysis," including Data Volume, Timeliness, Influence, and Serviceability. Not all dimensions from ISO/IEC 25012[14] are applicable in this context.

4.2.1 Data Volume. Data volume serves as the foundation of all data dimensions. Data volume refers to the number of security vulnerabilities in a vulnerability database, including the total and unique data.

Definition (Total Data). The total data volume of a vulnerability database is measured by the number of vulnerabilities it has collected as of Jan 1, 2021. Being the oldest data dimension, this attribute is crucial for any database.

Analysis Table 1 presents the data volume dimension results. NVD is the database with the largest amount of data, followed by CNNVD, which both have over 150,000 vulnerabilities. However, our measurements indicate that 144,096 vulnerabilities in both databases are identical CVE IDs, revealing a high degree of overlap between NVD and CNNVD. In contrast, the data in RNVD and ST is limited, possibly due to RNVD being newly established and ST no longer being maintained, resulting in data loss. Moreover, other databases have a certain bias and do not uniformly accept all vulnerabilities, instead selecting or requiring certain conditions to be met before accepting vulnerabilities.

Definition (Unique Data). The term "unique data" refers to the vulnerabilities exclusive to database A, compared to the CVE as of Jan 1, 2021. This unique data highlights the specific contributions of database A.

Analysis The unique vulnerability data is complementary to other vulnerability data. We analyzed the reasons for its generation. Firstly, some vulnerabilities may be ignored by CVE or not meet its criteria. Secondly, some databases have unique data due to regional issues, such as SeeBug, whose unique data primarily consists of Chinese software products not present in CVE. Finally, some databases do not have a complete mapping relationship with CVE IDs.

Additionally, Figure 3 provides insight into the change in the proportion of CVEs in various vulnerability databases over time. Notably, all databases, except EDB, seem to be converging towards CVE, which we refer to as "CVE assimilation." This suggests that the influence of CVE is gradually increasing.

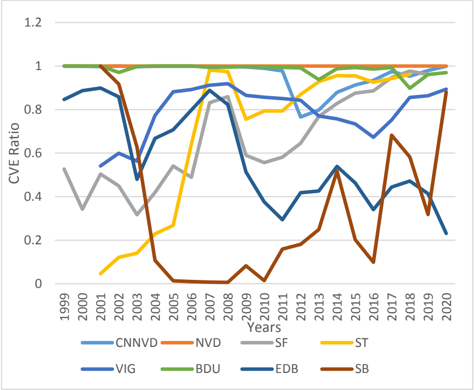

Figure 3: The CVE ratio over time.

4.2.2 Timeliness. **Definition (Timeliness).** Timeliness refers to the attribute of security vulnerability data regarding the time it was made public. It indicates the speed at which a database releases information about vulnerabilities.

We use NVD's release time as a benchmark to compare the timeliness of other vulnerability databases. The relative difference between a database's release time and NVD's is calculated. A positive value indicates that the database released its data after NVD, while a negative value means it was released before NVD. The published field in the vulnerability database records the availability of the security vulnerability. Before measuring the timeliness, we conduct compliance checks on the release times of each database to ensure their validity. For instance, a security vulnerability can't be released before 1988 or after Jan 1, 2021, which is our cutoff point. Any vulnerabilities published beyond this date are considered invalid.

Analysis SF has the quickest release time due to its vulnerability exposure policy, but SeeBug has a slower release time as it relies on a diverse group of white hat hackers and only releases information after verifying its relevance. In contrast, NVD takes more time to release vulnerabilities due to its thorough verification process to ensure information accuracy. Other vulnerability databases publish vulnerabilities at different times after the release of SF.

4.2.3 Influence. **Definition (Influence).** We assess the Influence of vulnerability databases by analyzing their reference links. The security vulnerability database's influence dimensions can be counted based on the reference links of different vulnerability information release sources.

Analysis We evaluated the reference value of different security vulnerability databases by analyzing the Top 5 sources with the most significant reference links for each database. These sources were mainly from major vendors such as Oracle or other vulnerability databases. The results of the influence evaluation are shown in Figure 4 after statistical analysis and normalization were performed.

Compared to CNNVD, SeeBug, and NVD, SecurityFocus (SF) has a greater impact due to its strict vulnerability exposure policy. SF is frequently the first to report vulnerabilities and restricts its disclosure time. If a company fails to respond to a vulnerability report

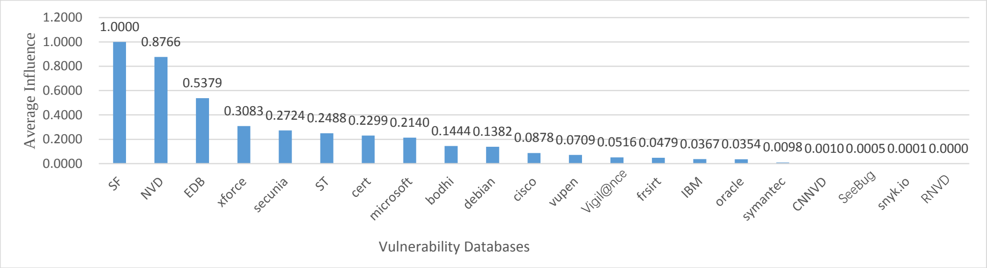

Figure 4: The influence of different vulnerability databases.

within a week, the reporter can directly disclose the vulnerability to SF. This one-week grace period gives SF an advantage in disclosing vulnerabilities. In comparison, the CVE has a 45-day waiting period prior to disclosing a vulnerability, while CNNVD only reveals vulnerabilities after a patch is issued. SF's comprehensive exposure policy, including original emails and summarized vulnerability information, sets it apart from the limited exposure policy of CVE and attracts both white and black-hat vulnerability researchers. The influence of databases such as CNNVD, RNVD, and SeeBug is minimal. This may be due to their leaky exposure policies, as well as their geographical attributes. For example, SeeBug focuses on Chinese software vulnerabilities not included in CVE. These findings indicate that a strong vulnerability exposure policy and a quick information exchange platform can garner greater attention and recognition.

Table 4: Service function list of different vulnerability databases.

DBs	CNNVD	RNVD	NVD	EDB	SF	ST	SeeBug	Vigil@nce
Subscription	×	○	○	×	×	×	×	○
Query	○	○	○	○	○	×	○	○
Weekly Rep	○	×	×	×	×	×	×	×
Monthly Rep	○	×	×	×	×	×	×	×
Type Ana	○	○	○	×	×	×	×	×
Threat Ana	○	○	○	×	×	×	×	×
VendorTop-N	○	○	×	×	×	×	×	×
ProductTop-N	○	○	×	×	×	×	×	×
Bulletin	○	×	×	×	×	×	×	×
Tendency Ana	○	○	○	×	×	×	×	×

4.2.4 Serviceability. **Definition (Serviceability).** Serviceability encompasses the process of summarizing and analyzing large amounts of data and is the ultimate objective of vulnerability databases. We have established subattributes, including **Subscription, Query, Weekly Report, Monthly Report, Vulnerability Type Analysis, Vulnerability Threat Analysis, Vendor Top-N, Product Top-N, Tendency, and Official Bulletin**. The serviceability provided by different vulnerability databases was evaluated based on these subattributes.

Analysis Vulnerability databases with effective data services, such as query functions and support for security vulnerability research, provide convenient access. Our survey results on the availability of these services are presented in Table 4.

The CNNVD has a favorable preference, and government-built vulnerability databases tend to have higher standardization than others. This aligns with the Compliance dimension. However, most vulnerability databases seem to have a limited emphasis on service functionality. Databases that only provide vulnerability reports lack analysis of growth trends, types, and other important information. Serviceability, on the other hand, offers a deeper understanding of potential vulnerabilities and helps in their evaluation.

5 THE COMPREHENSIVE ASSESSMENT MODEL

Using a comprehensive quality model, CQASVD applied Multi-Criteria Decision-Making (MCDM) methods such as AHP, entropy-based, and deviation maximization to carry out a comprehensive quality assessment of various vulnerability databases. The results of the evaluation can be found in Table 5.

Analytic Hierarchy Process. The Analytic Hierarchy Process (AHP), first introduced by Saaty[19], is a method for solving multi-objective decision-making problems as a system. By solving the judgment matrix eigenvector calculated for each element of each level on a hierarchy, the method prioritizes the weight of an element. Then, the method of weighted sum hierarchy merges various alternative solutions to the total target of the final weight. The final weight of the largest is the optimal program.

Entropy-Based method. Entropy was first developed by Shannon [20] and is used to measure the degree of disorder in a system. According to this theory, the natural trend is for systems to move toward disorder and increasing Entropy without external work. The entropy-Based method measures the degree of order of the system according to the amount of information. In the multi-object and multi-index systems, the weight value of different indexes is determined by the amount of information provided by each index in different systems. The entropy-Based method in MADM is objective and based on the amount of information. The higher the weight, the more important an attribute or database is.

Deviation Maximization method. Wang et al.[22] first proposed the objective weighting method based on maximizing deviation. This method assigns a greater weight coefficient to a data dimension when the difference between the data dimension values obtained by different schemes is larger, indicating that the data dimension is more critical to the decision-making of the scheme. Conversely, when the difference between the data dimension values obtained by different schemes is smaller, a smaller weight coefficient is assigned, indicating that the data dimension is less critical to the decision-making of the scheme.

Table 1 provides the measured values of different vulnerability databases in different data dimensions. In this section, we first normalize the data provided in Table 1, and then employ AHP, entropy-based, and maximizing deviation methods to assess vulnerability database data quality comprehensively.

Table 5: The results of the comprehensive evaluation model under different MADM methods.

Methods	NVD	CNNVD	RNVD	SF	ST	SeeBug	EDB	Vigil@nce
AHP	**0.1578**	0.1491	0.149	0.1515	0.0822	0.1178	0.0934	0.0992
Entropy	**0.098**	0.0822	0.0467	0.0939	0.0363	0.0385	0.0572	0.0402
Deviation Maximization	**0.7052**	0.6967	0.5369	0.6801	0.385	0.3299	0.4787	0.4484

The results displayed in Table 5 show that NVD, SecurityFocus, and CNNVD are very comprehensive databases and NVD is the best vulnerability database, as it provides the most detailed analysis of security vulnerabilities and keeps track of data validity. Despite some issues in the Patch and Exploit field, NVD remains a top choice. Meanwhile, although the SF database does not verify vulnerabilities, its strict exposure policy, comprehensive information, and strong impact make it a crucial resource for security vulnerability research. The CNNVD also performs well due to its completely structured information description and excellent services. While the EDB provides valuable PoC, it needs to improve its information structure. The RNVD has significant potential due to its fully structured data representation. The SeeBug vulnerability database still faces challenges with vulnerability collection, which need improvement. Vigil@nce, a commercial vulnerability database, is more focused on vulnerability repair.

6 DISCUSSION

6.1 The Comprehensive Quality Model

However, not all dimensions based on ISO/IEC 25012[14] are suitable for vulnerability databases. Other dimensions are not appropriate for the following reasons.

The CQASVD evaluation differs from Dong et al. [13] as we aim to assess the comprehensive quality of commonly used and credible vulnerability databases from a neutral perspective, encompassing factors beyond consistency, portability, and recoverability. The precision dimension was not considered as software version reports often contain numbers and letters. A vulnerability database should accurately convey information without causing ambiguity due to software abbreviations, which we measure through accuracy rather than understandability. Our analysis did not evaluate efficiency, as all vulnerability database data is stored in text format, and currentness, which refers to the data's relevance to the current situation, is inherent in the context of security vulnerability databases. Traceability is a key factor in monitoring vulnerabilities and updates reported by various sources. However, as evaluating and updating vulnerability information is standard practice in these databases, traceability is not considered a measurement dimension.

6.2 Some suggestions

Our comprehensive quality assessment analysis of various vulnerability databases reveals varying performance levels. These insights can be valuable in enhancing the quality of vulnerability databases.

Firstly, vulnerability databases should accept basic field information, such as a fixed field set containing crucial information about the vulnerability. Secondly, automated tools can be developed to efficiently review and verify vulnerabilities, thereby standardizing the processing of security vulnerabilities and addressing maintenance issues in vulnerability databases. Lastly, it is important to enhance the attention paid to patches. While the developer is responsible for patching vulnerabilities, the vulnerability database must be monitored and kept updated with information on any unpatched vulnerabilities.

7 RELATED WORK

Data quality is the degree to which data meets explicit and implicit requirements when used under specified conditions. Many scholars have proposed frameworks for data quality assessment from various perspectives in their research.

Chen et al.[11] have studied the data quality of public health information systems. They divided the dimensions of data quality into different attributes from three perspectives: the data, the intended use of the data, and the data collection process, and made a comprehensive evaluation of data quality based on these attributes.

Similarly, Laranjeiro et al.[16] have examined data quality evaluation attributes in recent years and have classified and mapped poor data's data quality issues to relevant dimensions. Lakshen et al.[15] have primarily focused on comprehensively describing the challenges of significant data quality and the functions required for big data processing. Lastly, Taleb et al.[21] have investigated, classified, and discussed the latest data quality assessment work. Croft[12] conducted a study on the data quality of software vulnerability datasets.

Zaveri et al.[23] conducted a comprehensive review on the evaluation of linked data quality. They proposed a comprehensive list of 18 quality dimensions and 69 metrics to aid in data evaluation. Similarly, Li et al.[9] have highlighted the characteristics of big data such as high volume, velocity, diversity, and value, and have analyzed the associated challenges that come with it.

8 CONCLUSION

This paper introduces CQASVD, a novel comprehensive quality assessment system for vulnerability databases. The system comprises two models: the comprehensive data quality model, which assesses the data quality of various vulnerability databases, and the comprehensive evaluation model, which evaluates their comprehensive performance. Using CQASVD, we evaluated several vulnerability databases, including CNNVD, RNVD, SeeBug, Exploit DB, SecurityFocus, SecurityTracker, and Vigil@nce. The results showed that NVD, SecurityFocus, and CNNVD are highly comprehensive databases. And NVD still leads as the best choice. SecurityFocus strongly impacts other vulnerability databases, while CNNVD is consistent with NVD and attains a high ranking to some degree. Our analysis highlight some of the challenges encountered in the operation and maintenance of vulnerability databases. In future work, we aim to refine the data quality model and employ more MCDM methods for better evaluating different databases.

REFERENCES

[1] 2020. BDU. https://bdu.fstec.ru/.

[2] 2020. CNNVD. http://www.cnnvd.org.cn/.

[3] 2020. EDB. https://www.exploit-db.com/.

[4] 2020. NVD. https://nvd.nist.gov/.

[5] 2020. SecurityFocus. https://www.securityfocus.com/.

[6] 2020. SecurityTracker. https://www.securitytracker.com/.

[7] 2020. SeeBug. https://www.seebug.org/.

[8] 2020. Vigli@nce. https://vigilance.fr/.

[9] Li Cai and Yangyong Zhu. 2015. The challenges of data quality and data quality assessment in the big data era. *Data science journal* 14 (2015).

[10] Oscar Chaparro, Jing Lu, Fiorella Zampetti, Laura Moreno, Massimiliano Di Penta, Andrian Marcus, Gabriele Bavota, and Vincent Ng. 2017. Detecting missing information in bug descriptions. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. 396–407.

[11] Hong Chen, David Hailey, Ning Wang, and Ping Yu. 2014. A review of data quality assessment methods for public health information systems. *International journal of environmental research and public health* 11, 5 (2014), 5170–5207.

[12] Roland Croft, M Ali Babar, and Mehdi Kholoosi. 2023. Data Quality for Software Vulnerability Datasets. *arXiv preprint arXiv:2301.05456* (2023).

[13] Ying Dong, Wenbo Guo, Yueqi Chen, Xinyu Xing, Yuqing Zhang, and Gang Wang. 2019. Towards the detection of inconsistencies in public security vulnerability reports. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 869–885.

[14] International Organization for Standardization/International Electrotechnical Commission et al. 2008. Software engineering-Software product Quality Requirements and Evaluation (SQuaRe) Data quality model. *ISO/IEC* 25012 (2008), 1–13.

[15] Guma Abdulkhader Lakshen, Sanja Vraneš, and Valentina Janev. 2016. Big data and quality: A literature review. In *2016 24th telecommunications forum (TELFOR)*. IEEE, 1–4.

[16] Nuno Laranjeiro, Seyma Nur Soydemir, and Jorge Bernardino. 2015. A survey on data quality: classifying poor data. In *2015 IEEE 21st Pacific rim international symposium on dependable computing (PRDC)*. IEEE, 179–188.

[17] Antonio Nappa, Richard Johnson, Leyla Bilge, Juan Caballero, and Tudor Dumitras. 2015. The attack of the clones: A study of the impact of shared code on vulnerability patching. In *2015 IEEE symposium on security and privacy*. IEEE, 692–708.

[18] Viet Hung Nguyen and Fabio Massacci. 2013. The (un) reliability of nvd vulnerable versions data: An empirical experiment on google chrome vulnerabilities. In *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security*. 493–498.

[19] Thomas L Saaty. 1977. A scaling method for priorities in hierarchical structures. *Journal of mathematical psychology* 15, 3 (1977), 234–281.

[20] Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.

[21] Ikbal Taleb, Mohamed Adel Serhani, and Rachida Dssouli. 2018. Big data quality: A survey. In *2018 IEEE International Congress on Big Data (BigData Congress)*. IEEE, 166–173.

[22] Wang Yingming. 1997. Using the method of maximizing deviation to make decision for multiindices. *Journal of Systems Engineering and Electronics* 8, 3 (1997), 21–26.

[23] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Soeren Auer. 2016. Quality assessment for linked data: A survey. *Semantic Web* 7, 1 (2016), 63–93.