# A Machine Learning Approach to Dataset Imputation for Software Vulnerabilities

Shahin Rostami[1,2], Agnieszka Kleszcz[3], Daniel Dimanov[1], and Vasilios Katos[1]

[1] Bournemouth University, Poole, UK {srostami,vkatos}@bournemouth.ac.uk
[2] Polyra Limited, Bournemouth, UK shahin@polyra.com
[3] AGH University of Science and Technology, Cracow, Lesser Poland District, Poland
akleszcz@agh.edu.pl

**Abstract** This paper proposes a supervised machine learning approach for the imputation of missing categorical values from the majority of samples in a dataset. Twelve models have been designed that are able to predict nine of the twelve ATT&CK tactic categories using only one feature, namely the Common Attack Pattern Enumeration and Classification (CAPEC). The proposed method has been evaluated on a 867 sample unseen test set with classification accuracy in the range of 99.88%-100%. Using these models, a more complete dataset has been generated with no missing values for the ATT&CK tactic feature.

**Keywords:** Cyber Security · Vulnerability · Mitre ATT&CK · Machine Learning · Dataset · Imputation

## 1 Introduction

Software vulnerabilities are a representative cause of security policy violations in computer systems. The omnipresent nature of vulnerabilities as evidenced by the constantly increasing number of discovered vulnerabilities per year has triggered significant efforts in their study. The importance and impact of vulnerabilities on practical computer security have led to the development of vulnerability management frameworks and analysis approaches, see for example the vulnerability lifecycle [1]. At the same time, the progress in the field of machine learning and its applications in a number of domains has sparked a body of research on analytics for cybersecurity related problems.

Unsurprisingly, datasets are a key ingredient to such thread of research and in cybersecurity there are many readily available datasets published by the research and academic community as well as by the cybersecurity industry. In cybersecurity we could distinguish two types of dataset based on the way they are generated; one type consists of datasets that are synthesised from emulated or simulated data in order to enable the research community to study a particular computing environment. The other type consists of datasets that are generated from actual security incidents, i.e. real-world data. Honeynet and honeypot data are also included in this category as they capture events from real attackers, but their actions do not have an actual impact on the target infrastructure and

business. Interestingly, vulnerability related data fall under the second category. This makes the collection of the relevant data challenging, as it requires one to invest on a systematic effort to triage, evaluate, consolidate and catalogue the vulnerability data from different sources in order to develop a meaningful dataset. Acknowledging the limitations and challenges the development of a complete vulnerability dataset entails, this research proposes an approach to bridge the gap between the two dataset types by enriching a vulnerability dataset with synthetic data in a way that it will enable further study of vulnerabilities.

### 1.1 Motivation

Obtaining and maintaining complete and high quality datasets is not a trivial task. Despite the wealth and availability of disparate datasets in many domains, information sharing of cybersecurity related data displays certain nuances; researchers may have a financial motivation not to share vulnerability information and particularly any associated with zero-day vulnerabilities. This is mitigated to some extent by bug bounty programmes and responsible disclosure policies. Organisations who deploy and use the software to deliver their business models on the other hand may not be forthcoming in disclosing attack and vulnerability information as this may result in a higher risk of exposure. Finally, third parties who have developed business models on the commercialisation of threat information sharing are understandably reluctant to support the wider community with freely available data and prefer to make these available to their premium customers.

All the above contribute to a tessellated landscape of cybersecurity datasets that can be of varying reliability and trustworthiness as well as being incomplete. Furthermore, despite the ongoing attempts to standardise the expression of threat data through a number of taxonomies, the actual datasets end up having conflicting or contradictory values.

**Contribution.** This paper focusses on the incompleteness aspect of a vulnerabilities dataset by proposing a supervised machine learning approach for the imputation of missing categorical values from the majority of samples. The feature selected for the imputation is the ATT&CK tactic, which is typically used to communicate the high level modus operandi of an attacker. Twelve models have been designed that are able to predict nine of the twelve ATT&CK tactic categories using only one feature, namely the Common Attack Pattern Enumeration and Classification (CAPEC). This has been evaluated on a 867 sample test set with classification accuracy in the range of 99.88%-100%. Using these models, a more complete dataset has been generated with no missing values for the ATT&CK tactic feature.

## 2 The emerging ecosystem of software vulnerabilities

Vulnerabilities constitute a key element of ICT systems security as they enable both threat actors and defenders to realise their respective and competing agen-

das; an attacker would exploit the vulnerability in order to succeed in system compromise, whereas a defender would use the knowledge to conduct, inform, and eventually establish an effective and practical risk management plan. As vulnerabilities contribute to actionable cyber threat intelligence, they also inherit the properties and quality requirements of such type of information, such as relevance, timeliness, accuracy, completeness, and ingestibility [2]. Moreover the description of vulnerabilities has been fairly standardised, with the Common Vulnerabilities and Exposures (CVE) [3] programme being the most popular convention for cataloguing and classifying vulnerabilities. The CVE catalogue has been enriched with further initiatives such as the CVSS scoring system which attaches a quantitative measure of the severity of a particular vulnerability, the Common Attack Pattern Enumeration and Classification (CAPEC) [5] programme that associates vulnerabilities with attacks, and the more generic Common Weakness Enumeration (CWE) [4] that attempts to represent the software weaknesses through a standardisation language. A vulnerability that enters the aforementioned ecosystem has a minimum requirement of a CVE identification (CVE-id), whereas any other information could be optional. Although there is in principle an "authoritative" database with the CVE-ids, ensuring that these ids are unique and refer to vulnerabilities in an non-arbitrary manner, all other descriptors are not necessarily complete or correct. In fact, it was found in [6] that CVSS scores for the same CVE-id can differ significantly between different versions or databases.

Apart from the generic quality criteria that vulnerabilities inherit form being actionable cyber threat intelligence items, they also have their own, esoteric ones. The authors in [7] list two main categories that a vulnerabilities database (or dataset) should cover, namely *information coverage* and *capabilities*. In terms of the former, the evaluation criteria cover the *scope, impact & risk, resolution, vendor, products, exploit, categorisation*, and *relations*. Regarding the capabilities of the database, the authors highlight the supported *standards*, the existence and prevalence of a *community* adopting and supporting the data, the *interfacing* capabilities, and *freshness* of the contained data.

Vulnerabilities are also observed through their so-called vulnerability lifecycle [8]. The lifecycle introduces a chronological contextualisation to the vulnerability by identifying significant milestones and events that define risk-transitioning boundaries. More specifically, upon the discovery of a vulnerability, the associated risk follows an upward trend which spikes when a practical exploit is created; if this happens prior to the notification of the software vendor the risk reaches the highest peak, as the exploit will be considered a *zero-day*. The researcher who discovered the vulnerability may choose to notify the vendor following a responsible disclosure practice, or publicise its details. Bug bounty programmes attempt to regulate and streamline the vulnerability discovery and reporting process through financial incentive schemes. It should be evident that for each of the aforementioned events the risk will be affected. As such, timing - and timeliness - are significant and influencing factors.

The vulnerabilities may also be studied through an organisational and geopolitical perspective. In [9] the authors examine whether there are differences between different China-based organisations in respect of their status or sector (established, public sector, education, or startup), revealing that startups experienced the biggest challenges. Such a study was possible by employing publicly available vulnerability data.

The efforts to increase both the understanding and effective sharing of vulnerabilities are also reflected through the emergence of frameworks and tools such as STIX and the ATT&CK framework. In STIX, the specification language for structured cyber threat intelligence sharing, vulnerabilities are expressed through a dedicated and specific object type. The ATT&CK framework is a curated knowledge-base of adversarial *techniques* and *tactics*. As these have recently gained popularity, not all published vulnerabilities have been mapped or assigned to the above schemes. Enriching the datasets with these dimensions is expected to generate considerable added value.

## 3 The ENISA vulnerabilities dataset

When constructing a dataset from multiple sources it is anticipated that this would inevitably lead to having empty values, as the different data sources do not necessarily overlap horizontally. The vulnerability dataset contains missing values due to the missing data from the source database but also due to the operation of joining the different sources. In December 2019, the European Union Agency for Cybersecurity (ENISA) published a report entitled "State of Vulnerabilities 2018/2019: Analysis of Events in the life of Vulnerabilities" [6]. This dataset covers the period of vulnerabilities published between January 1st 2018 to August 31st (Q1 – Q3) 2019. The vulnerabilities were collected and hosted in the compiled dataset until the cut-off date of September the 30th. The data is organised into a two-dimensional tabular structure in the shape of (27471 rows $\times$ 59 columns). Out of the 59 columns, those containing the vulnerability id, CVSS scores (both versions), Common Weakness Enumeration (CWE), and the number of exploits had completely filled values, although the number of exploits had the value of 0 on over 90% of the vulnerabilities. The noteworthy columns of missing values are CAPEC (77% completed), ATT&CK techniques and tactics (approx. 29% complete), and price information (approx 12% complete). In terms of the Common Platform Enumeration (CPE), the vendor and product information was complete at 84%, but the platform information had low completion, with only 8.6% of the values being populated. The smallest measure of completeness was observed in the sector information with only 0.5% completion. In terms of absolute numbers, this amounted to 137 vulnerabilities annotated with sector information. Although this allowed the execution of some rudimentary statistical tests, this is was not considered adequate for more advanced research techniques using machine learning.

The dataset combines different open sources such as the National Vulnerability Database (NVD), Common Weakness Scoring System (CWSS), Common

Vulnerabilities and Exposures (CVE), Shodan, Zerodium, and so forth. Table 1 presents data sources. The dataset was made publicly available[4] together with the associated Jupyter Notebooks in order to allow the research community to scrutinise the findings contained in the report, but also to enable further research.

Table 1: Data sources characteristic[6]

| **Source Type** | **Data Type** | **Description** |
|---|---|---|
| NVD database | CVE data | The NVD is the U.S. government repository of standards-based vulnerability management data. The NVD includes databases of security checklist references, security-related software flaws, misconfigurations, product names, and impact metrics [5]. |
| ATT&CK | Attacker's patterns (techniques & tactics) | MITRE ATT&CK™ is a globally-accessible knowledge base of adversary tactics and techniques based on real-world observations [6]. |
| Shodan | Number of exploits | Database of internet connected devices (e.g. webcams, routers, servers, etc.) acquiring from various HTTP/HTTPS - port 80, 8080, 443, 8443) [7]. |
| Exploit database | Non-CVE data | Contains information on public exploits and corresponding vulnerable software. The collection of exploits is acquired from direct submissions, mailing lists and other public sources [8]. |
| CVE details | CVE data | The database containing details of individual publicly known cybersecurity vulnerabilities including an identification number, a description, and at least one public reference [9]. |

[4] https://github.com/enisaeu/vuln-report
[5] bibitemnvdhttps://nvd.nist.gov
[6] https://attack.mitre.org
[7] https://www.shodan.io
[8] https://www.exploit-db.com/about-exploit-db
[9] https://cve.mitre.org

Table 1: Data sources characteristic[6]

| Source Type | Data Type | Description |
|---|---|---|
| Zero-Day Initiative | CVE and non-CVE | Encourages reporting of zero-day vulnerabilities privately to affected vendors by financially rewarding researchers (a vendor-agnostic bug bounty program). No technical details on individual vulnerabilities are made public until after vendor released patches. ZDI do not resell or redistribute the vulnerabilities [10]. |
| ThreatConnect | Number of incidents related to CVE | Automated threat intelligence for Intel systems [11]. |
| VulDB | Exploit prices and software categories | Vulnerability database documenting and explaining security vulnerabilities and exploits [12]. |
| US CERT | Industry sector | The US Department for Homeland Security's Cybersecurity and Infrastructure Security Agency (CISA) aims to enhance the security, resiliency, and reliability of the USA's cybersecurity and communications infrastructure [13]. |
| Zerodium | Bug bounty exploit prices | A zero-day acquisition platform. Founded by cyber security experts with experience in advanced vulnerability research [14]. |

## 4 Dataset imputation through machine learning

Since the influential publications of Rubin many decades ago, e.g. [12,11], there has been increasing awareness of the drawbacks associated with analyses conducted on datasets with missing values. This is prevalent in many fields of study, particularly medical (clinical) datasets which can often be missing values for over half of the samples available [15].

Classical imputation methods often relied on the use of measures of central tendency of available data to populate the missing values, e.g. the arithmetic

[10] https://www.zerodayinitiative.com
[11] https://threatconnect.com
[12] https://vuldb.com
[13] https://www.us-cert.gov
[14] https://zerodium.com

mean, mode, and median. However, these methods are now considered to be ineffective for computing candidates for the population of missing data and are more likely to reduce the accuracy and integrity of a dataset [10], e.g. in the case of heteroskedasticity. Hot-deck imputation, an approach which uses randomly selected *similar* records to impute missing values, performs poorly when the majority of samples contain missing values and are outperformed by other approaches [18].

More recent studies suggest that artificial neural networks, particularly multilayer perceptrons [16] and autoencoders [17], can outperform these classical methods, including regression and hot-deck, for the imputation of categorical variables. Artificial neural network methods have also been shown to outperform Expectation-Maximisation techniques in the presence of non-linear relationships between sample variables [19].

Following the suggestions in the literature a machine learning approach, i.e. artificial neural networks, will be used for the imputation of missing categorical values. This experiment aims to estimate categorical variables where there are missing values in the *tactics* feature of the ENISA vulnerabilities dataset. This will be achieved by detecting patterns in the sub-components of the CAPEC and their mappings to one or many tactic categorical values. This will be treated as multiple sub-problems, whereby each tactic can be considered a flag on a binary string where the binary value is determined by a binary classifier. The dataset consists of 27471 samples, where 19404 of these samples have no value for the tactic feature. Each sample can be labelled with multiple unique tactic categories from the following list:

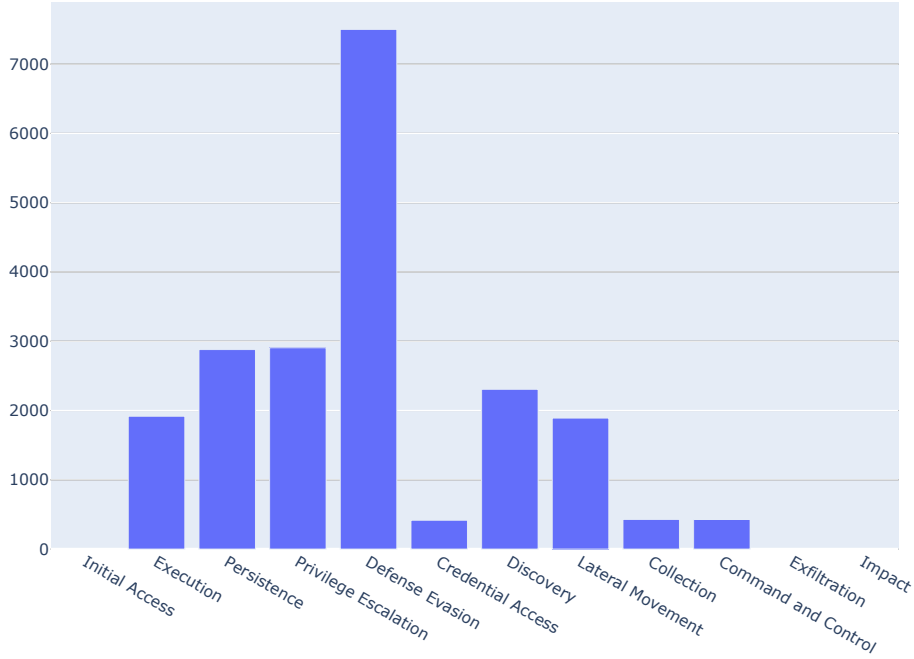
- Initial Access
- Execution
- Persistence
- Privilege Escalation
- Defense Evasion
- Credential Access
- Discovery
- Lateral Movement
- Collection
- Command and Control
- Exfiltration
- Impact

Figure 1 illustrates the distribution of tactic category assignments where it can be seen that three of these twelve categories are not represented at all. Without examples for all twelve categories it is expected that a dataset populated through imputation will also not represent these entirely missing categories.

The CAPEC cannot be used directly and therefore it must first be preprocessed and encoded. For reproducibility, this consisted of the following steps:

1. All CAPEC IDs were obtained using the Domains of Attack resource from Mitre (https://capec.mitre.org/data/definitions/3000.html).

**Figure 1.** The distribution of ATT&CK tactic labels within the original dataset.




2. The CAPEC IDs are then used to create a truth table with all elements initialised to false.
3. Using the CAPEC feature in the ENISA Vulnerabilities dataset, the corresponding CAPEC ID in the truth table is changed to true.

The tactic feature required similar encoding as it exists as a comma separated variable in the ENISA vulnerabilities dataset:

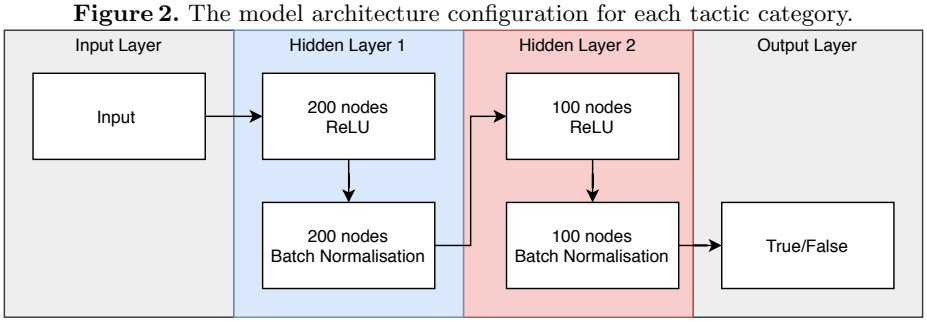
1. The tactics were obtained using the ATT&CK Matrix resource from Mitre (https://attack.mitre.org/)
2. The tactics are then used to createa truth table with all elements initialised to false.
3. Using the tactic feature in the ENISA Vulnerabilities dataset, the corresponding tactic in the truth table is changed to true.

These two truth tables were used to design individual models per tactic using supervised machine learning. The inputs for all the models were to be the same, the CAPEC truth table, whereas the output for each of the twelve models could be either True or False.

Each model was designed to have the same architecture presented in Figure 2, a 2 hidden layer feedforward neural network. The hidden layers used ReLu nonlinear activation functions and batch normalisation, with the first hidden

layer consisting of 200 nodes and the second one consisting of 100. We selected this architecture based on successful examples on tabular data in the literature [21] [20].

**Figure 2.** The model architecture configuration for each tactic category.




The optimiser employed during the supervised learning process was the Adam algorithm configured with 0.9 and 0.99 respectively as the beta coefficients used for computing running averages of gradient and its square, a weight decay rate of $1e-2$, and a learning rate of $1e-3$. Adam was selected because of its performance and memory advantage over other optimisation algorithms, especially for multivariate data [22].

All twelve models were trained on a 6500 sample subset of the 8067 samples for which the tactic categorisations were present. A further 700 of these samples were used for validation, with the final 867 samples reserved for testing.

Each model was trained for 5 epochs before achieving at least 99.88% accuracy, indicating the likelihood of a prominent pattern in the mappings of CAPECs to tactics.

On the unseen data reserved for the test set the models predicting Persistence, Privilege Escalation, Defence Evasion, Credential Access, and Collection achieved 99.88% accuracy, with the remaining classifiers achieved 100%.

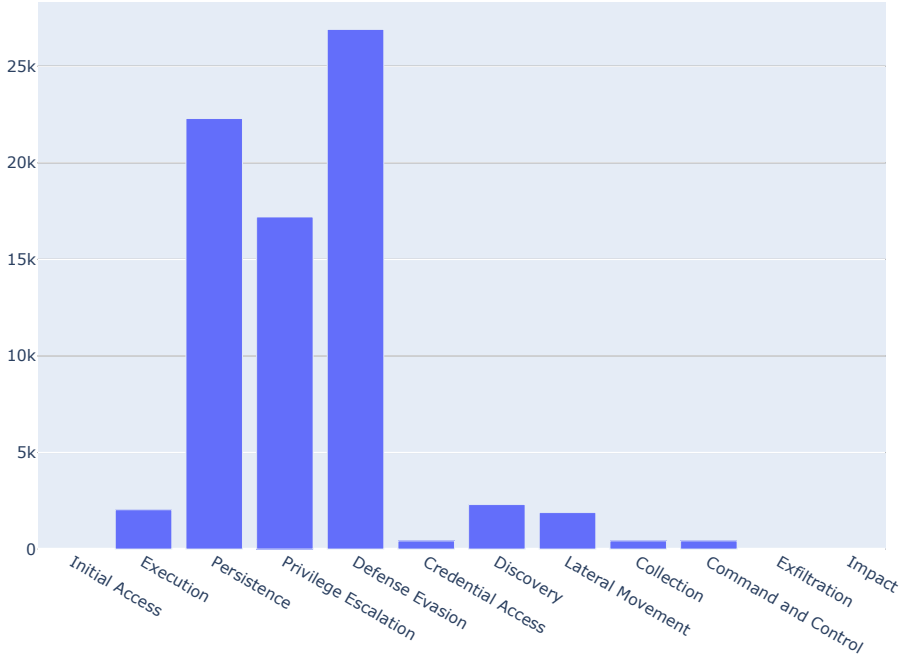
The models were then used to collectively predict the tactics for the entire dataset, including those 19404 samples which had missing values. Figure 3 illustrates the distribution of tactic category assignments in the dataset with imputation, where it can be seen that as expected the same three categories are still not represented at all.

## 5 Discussion

Whilst from a machine learning perspective the results are completely satisfactory, a number of limitations need to be acknowledged.

First, the ATT&CK tactics which are closely coupled to the concept of the Cyber Kill Chain (CKC) [23], inherit the limitations and disadvantages of such

**Figure 3.** The distribution of ATT&CK tactic labels within the imputed dataset.




taxonomy. Specifically, the CKC phases were mainly intended to help communicate the main stages of a cyber attack as a sequence of events, similar to the description of an attack vector. Mapping a particular attack action to the respective kill chain phase carries a degree of subjectivity. This can be easily evidenced by studying the structure and approach of the ATT&CK framework, where the techniques do not have an exclusive membership under the tactics. In fact, there are some techniques that can appear in as many as four different tactics. As such, when investigating a particular cyber incident, the placement of an identified technique is a task for the security analyst. Therefore, the proposed machine learning method has not taken into account such implicit knowledge. This is a much more complex problem and certainly deserves a separate and dedicated research thread. A way forward for future research is to include more features and particularly the techniques which can appear having multiple values per vulnerability. This multiplicity will allow the application of machine learning techniques to identify the distances of the data points.

Second, there were differences between the original distribution and that of the imputed dataset. As the accuracy was high, we can conclude that the imputation revealed some interesting information on the tactics. Specifically, the original dataset showed that *Defence Evasion* was by far the most frequent attack, but in the imputed data *Persistence* and *Privilege Escalation* are now

comparable. Intuitively it is likely that this is correct as these three tactics can be seen to have a symbiotic relationship.

Finally, it should be noted that the dataset itself contains potentially contradictory data. This is an inherent problem when ingressing data from multiple sources. In addition, as noted in the vulnerability study, the transition to a revised vulnerability scoring system (namely from CVSS version 2 to version 3.x) resulted in discrepancies, despite the two scoring systems being obtained from the same and relatively "authoritative" database [6]. This could potentially be problematic when performing an imputation if it is not clear which of the competing features are "correct", as in this case the accuracy performance of the imputation algorithm may be irrelevant.

## 6 Conclusions and Future Work

In this paper we attempted to enrich a real-world dataset using supervised machine learning. We demonstrated that it is possible to completely fill a sparse column of the dataset and we selected the ATT&CK tactic feature to showcase this approach.

The significance of the results is two-fold. First, the high accuracy achieved showed the performance and feasibility of the proposed approach. Second, we demonstrated that it is possible to escape from the inherent limitation where only real-world data are available due to the nature of the problem, e.g. the study of vulnerabilities which cannot be created by completely synthesised data.

As a future direction, we will investigate the imputation of other features, including those that describe financial aspects of the vulnerabilities.

## Acknowledgement

This work has received funding from the European Union's Horizon 2020 research and innovation program under the grant agreement no 830943 (ECHO).

## References

1. Joh, H. and Malaiya, Y. A Framework for Software Security Risk Evaluation using the Vulnerability Lifecycle and CVSS Metrics, Proc. International Workshop on Risk and Trust in Extended Enterprises, pp. 430–434 (2010)
2. ENISA: Actionable Information for Security Incident Response. Heraklion, Greece (2015) https://doi.org/10.2824/38111
3. MITRE, Common Vulnerabilities and Exposures https://cve.mitre.org/. Last accessed 16 Feb 2020
4. MITRE, Common Weakness Enumeration https://cwe.mitre.org/. Last accessed 16 Feb 2020
5. MITRE, Common Attack Pattern Enumeration and Classification https://capec.mitre.org/. Last accessed 16 Feb 2020

6. ENISA: State of Vulnerabilities 2018/2019 - Analysis of Events in the life of Vulnerabilities, Heraklion, Greece, 2019. https://www.enisa.europa.eu/publications/technical-reports-on-cybersecurity-situation-the-state-of-cyber-security-vulnerabilities/at_download/fullReport
7. Kritikos, K., Magoutis, K., Papoutsakis, M., Ioannidis, S.: A survey on vulnerability assessment tools and databases for cloud-based web applications. Array **3–4**(100011), 1–21 (2019)
8. Arbaugh, W., Fithen, W., McHugh, J.: Windows of Vulnerability: A Case Study Analysis. IEEE Computer **3**(12), 52–59 (2000)
9. Huang, C., Liu, J., Fang, Y., Zuo, Z.: A study on Web security incidents in China by analyzing vulnerability disclosure platforms. Computers and Security **58**, 47–62 (2016)
10. Royston, Patrick. "Multiple imputation of missing values." The Stata Journal 4.3 (2004): 227-241.
11. Rubin, Donald B. Multiple imputation for nonresponse in surveys. Vol. 81. John Wiley Sons, 2004.
12. Rubin, Donald B. "Inference and missing data." Biometrika 63.3 (1976): 581-592.
13. Rubin, Donald B. "Multiple imputation after 18+ years." Journal of the American statistical Association 91.434 (1996): 473-489.
14. Rubin, Donald B., and Nathaniel Schenker. "Multiple imputation in health-are databases: An overview and some applications." Statistics in medicine 10.4 (1991): 585-598.
15. Clark, Taane G., and Douglas G. Altman. "Developing a prognostic model in the presence of missing data: an ovarian cancer case study." Journal of clinical epidemiology 56.1 (2003): 28-37.
16. Silva-Ramírez, Esther-Lydia, et al. "Missing value imputation on missing completely at random data using multilayer perceptrons." Neural Networks 24.1 (2011): 121-129.
17. Choudhury, Suvra Jyoti, and Nikhil R. Pal. "Imputation of missing data with neural networks for classification." Knowledge-Based Systems 182 (2019): 104838.
18. Wilmot, Chester G., and Shivaprasad Shivananjappa. "Comparison of Hot-deck and Neural-network Imputation." Transport survey quality and innovation (2003): 543-554.
19. Nelwamondo, Fulufhelo V., Shakir Mohamed, and Tshilidzi Marwala. "Missing data: A comparison of neural network and expectation maximization techniques." Current Science (2007): 1514-1521.
20. Guo, Cheng, and Felix Berkhahn. "Entity embeddings of categorical variables." arXiv preprint arXiv:1604.06737 (2016).
21. De Brébisson, Alexandre, et al. "Artificial neural networks applied to taxi destination prediction." Proceedings of the 2015th International Conference on ECML PKDD Discovery Challenge-Volume 1526. (2015).
22. Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." ICLR (2015).
23. Hutchins, E., Cloppert, M. and Amin,R. Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains. Bethesda, MD: Lockheed Martin Corporation (2010) https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/LM-White-Paper-Intel-Driven-Defense.pdf. Last accessed 16 Feb 2020