UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

LUCAS ROTTA KRÜGER

Data Quality in Machine Learning and Network Security: A Systematic Literature Review

Work presented in partial fulfillment
of the requirements for the degree of
Bachelor in Computer Science

Advisor: Prof. Dr. Lisandro Z. Granville
Coadvisor: Dr. Éder John Scheid

Porto Alegre
January 2025

ABSTRACT

Network security is a critical issue in the field of computer science, with data quality being one of its most important aspects. This work provides a systematic literature review of existing research that is directly relevant, tangentially related, or potentially applicable to analyzing the quality of network data for training machine learning models capable of sustaining computer networks. Key concepts related to Data Quality (DQ), Machine Learning (ML), and Network Security are introduced, along with a discussion on how DQ can be assessed and applied to ML or network security. The aim of this work is to identify, evaluate, and summarize the research relevant to a set of defined research questions. Understanding the technologies currently employed is essential for recognizing trends, benefits, and limitations, as well as identifying research gaps and establishing directions for future studies to guide scholars in this field.

Keywords: Data quality. Machine learning. Computer networks. Literary review.

Qualidade de Dados em Aprendizado de Máquina e Segurança de Rede: Uma Revisão Sistemática da Literatura

RESUMO

A segurança da rede é um problema crítico no campo da ciência da computação, sendo a qualidade dos dados um de seus aspectos mais importantes. Este trabalho apresenta uma revisão sistemática da literatura sobre pesquisas existentes que são diretamente relevantes, tangencialmente relacionadas ou potencialmente aplicáveis à análise da qualidade dos dados de rede para o treinamento de modelos de aprendizado de máquina capazes de sustentar redes de computadores. Os principais conceitos relacionados à qualidade dos dados (DQ), ao aprendizado de máquina (ML) e à segurança da rede são apresentados, juntamente com uma discussão sobre como a DQ pode ser avaliada e aplicada ao ML ou à segurança da rede. O objetivo deste trabalho é identificar, avaliar e resumir a pesquisa relevante para um conjunto de perguntas de pesquisa definidas. Compreender as tecnologias atualmente empregadas é essencial para reconhecer tendências, benefícios e limitações, bem como identificar lacunas de pesquisa e estabelecer direções para estudos futuros a fim de orientar os acadêmicos nesse campo.

Palavras-chave: Qualidade de Dados, Aprendizado de Máquina, Redes de Computadores, Revisão da Literatura.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Lisandro Z. Granville, for his invaluable guidance, encouragement, and support throughout my research and thesis writing process.

I am also particularly grateful to my co-advisor Dr. Éder John Scheid for his essential support, expertise, and close involvement in every step of the way.

I am also grateful to my colleagues and lab mates for their collaboration and insightful discussions.

Lastly, I would like to thank my family and friends for their unwavering support and understanding throughout this journey.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ML	Machine Learning
DQ	Data Quality
QA	Quality Assessment
SLR	Systematic Literature Review
RSLS	Reviews with Systematic Literature Searches

CONTENTS

1 INTRODUCTION

A significant challenge in the field of Machine Learning (ML) is the Data Quality (DQ) of datasets used in ML algorithms. The performance of these algorithms, in terms of both efficiency and accuracy, is directly influenced by the quality of the data employed during the training and testing phases. To address this issue, various metrics have been defined, including precision, timeliness, uniqueness, validity, consistency, and completeness.

Established metrics, though, are generally quite context agnostic, not accounting for the specific domain where the data is applied. This presents a challenge: identifying and understanding the tailored set of quality metrics required to quantitatively evaluate datasets available for public use, particularly in the domain of network security, which are used to develop ML models. To better understand the current state of research in this field, a systematic literature review is essential to consolidate existing knowledge and identify gaps or areas for improvement.

This work aims to conduct a systematic literature review to analyze the state of the art in defining and applying data quality metrics for network security datasets used in ML. The study will systematically map the existing research proposed in the literature and answer the following Research Questions (RQ):

- RQ1 - Is there a consensus on the metrics to analyze DQ?
- RQ2 - In which fields is DQ assessment the most prevalent?
- RQ3 - How DQ is monitored or considered in the ML research?
- RQ4 - Are there efforts to quantify DQ for ML in the network security context?

This will provide a foundation for advancing methodologies in assessing data quality in ML applications for network security and related domains. Thus, this work is organized in as follows. Chapter 2 presents the background on the concepts involved, Chapter 3 presents the methodology of the literature review, Chapter 4 describes selected approaches, Chapter 5 answers the RQs, and, lastly, Chapter 6 concludes the work.

2 BACKGROUND

In this section key concepts related to Data Quality (DQ), Machine learning (ML), and Network Security are presented. This section serves to introduce the terms and definitions essential to the mentioned areas of research.

2.1 Data Quality

For the remainder of this literature review, DQ will be mainly relevant when applied in ML and network research. Nevertheless, its concepts extend much beyond these areas, and as such, many of the traditional principles can be useful when considering how to tailor methods or metrics specific to an area of application, in this case, ML training data. The actual DQ assessment is often defined as "fitness for use" (WANG; STRONG, 1996). In other words, DQ is intrinsically related to how well it can be used by the consumer, and that is how it should ultimately be measured. General definitions though useful for when establishing the analyses protocol usually take a background to use specific evaluations (XU; ZHANG; SHI, 2022) (CAI; ZHU, 2015).

2.1.1 Measurements and Data Collection Issues, Concepts, and Definitions

In this subsection, we briefly review the definitions of errors from the data collection phase (MINING, 2006). Much of the issues that can compromise DQ can come from the data collection phase, where the resulting values might be missing, duplicated, inconsistent, or even included from outside of the scope of interest. All of these can be encompassed by the term **data collection error**. In addition, in cases where the measurement process results in data that differ from the true value, the data are said to contain **measurement errors**. Both types of error can be caused by systematic issues or random errors in the process.

Noise is a term that specifies a type of measurement error that encompasses some randomness. Often originating from the distortion of a value or the addition of extra objects, it is usually connected to data that has a spatial or temporal component, and when the noise is too high it can significantly compromise the overall pattern and, consequently, the conclusion that would originate from it. The elimination of noise, though it

is frequently non trivial and quite difficult to produce, has been amply studied and often produces significant DQ improvement (LI et al., 2016) (XIONG et al., 2006). Artifact is a similar concept to noise, though it refers to errors caused by a more deterministic phenomenon and consequently with a more consistent effect, such as streak in the same spot of multiple photographs, caused by a spot in the lens.

Precision is the closeness of repeated measurements (of the same quantity being measured), usually based on the standard deviation of a set of values. Accuracy is the closeness of measurements to the true values of the quantity being measured and may be based on precision. Outliers have characteristics outside of the common pattern of the rest of the data, though, similar to the definition of noise, outliers can be legitimate data objects or values, and their analyses can potentially lead to a more complete representation. Outliers can be caused either by collection errors or from the fact that not all attributes are applicable to all data objects; regardless, they should be taken into account during data analysis.

Inconsistent values are data values that, when translated to the supposed corresponded real object, are proven impossible, say a negative interval of time or mass of an object. Duplicate data are when two entries refer to the same data object. When dealing with such, deduplication, inconsistent values could be generated and must also be dealt with.

Timeliness is especially relevant when data represent a snapshot of a ongoing phenomenon, then data are only truly representative when it is new, and old data represent an environment that no longer exists. Relevance considers that data must contain fields that affect the hypothesis being tested under the penalty of reduced accuracy. Defining precisely which attributes are useful can also be hard to determine, since more harmful causes- such as sampling bias- can provide ratios that do not correspond to the truth of the phenomenon being studied. Overall knowledge of the conventions used when storing data is also fundamental to determining the correct ways to analyze it, e.g. empty fields being represented by a big negative number can easily compromise mean averages.

2.1.2 Traditional Dimensions

A quintessential slice of DQ work involves dimensions and metrics, as such a at least passing familiarity with it's most often relevant examples makes itself necessary to partake in any discussion in the topic of data quality.

Accuracy, although its definition can vary depending on the field in which it is used, in general can be reduced to the notion of the magnitude of an error, without much loss of meaning (HAEGEMANS; SNOECK; LEMAHIEU, 2016). Although varying it being an intrinsic characteristic of the data (PIPINO; LEE; WANG, 2002), it has the limitation of needing reliable reference data to produce meaningful value comparisons. For a more precise definition (SCANNAPIECO, 2006) Accuracy is defined as the closeness between a value v and a value v', considered as the correct representation of the real-life phenomenon that v aims to represent.

Completeness is usually defined as the extent to which the data are not missing. It is a contextual metric, that is, to judge it; is necessary to have previously defined what is a complete entry in the current experiment (PIPINO; LEE; WANG, 2002). Completeness can be generically defined as "the extent to which data are of sufficient breadth, depth, and scope for the task at hand" (SCANNAPIECO, 2006).

Consistency refers to the validity and integrity of data representing real-world entities, with the goal of detecting inconsistencies of conflicts in the data. In relational databases they may manifest within a single tuple, between different tuples in the same relation (table), and between tuples across different relations (FAN; GEERTS, 2022)

2.1.3 Data Quality Issues

Dirty data (KIM et al., 2003) Data are dirty if the user or application ends up with a wrong result or cannot derive a result due to certain inherent problems with the data. The sources of dirty data include data entry error by a human or computer system, data transmission error by a computer system, and even bugs in a data processing computer system. The basic data cleansing solution could involve the "Basic Sorted-Neighborhood Method" in which after concatenating two databases with suspected dirty data, relevant fields are selected and through the use of an error model, keys are created. In the sequence the data gets sorted and keys that are more similar than a chosen parameter highlight data that those fields should have been treated as one but got "corrupted" (HERNÁNDEZ; STOLFO, 1998).

Missing values are usually attributed to human error when processing data and machine error due to malfunctioning equipment, respondents' refusal to answer certain questions, dropout in studies, and merging of unrelated data (EMMANUEL et al., 2021). The treatment of missing data can be executed by means of: deletion of instances, re-

placement with estimated values, and imputation. Imputation itself can be implemented by means of mean, regression, K nearest neighbor, or ensemble-based approaches. Although many different approaches have been studied to solve missing data issues, it is important to note that the only suitable solution comes down to a virtuous design and good analyses (EMMANUEL et al., 2021).

Duplicate data problems are ever more common, with an ever increasing volume of data from many, distributed, and heterogeneous sources, such DQ issues are bound to happen. Duplicate data can be precisely defined by "Multiple yet different representations of the same real-world object" (NAUMAN; HERSCHEL, 2022), and can be generated from two types of sources, intra-source duplicates are a single entity inadvertently entered multiple times into the same database, usually appear when entities are entered without properly checking for existing representation, often from human error or (OCR) optical character recognition, frequently used for digitization of hand-written sources. Inter-source duplicates, on the other hand, often appear when integrating multiple data sources that might have a different representation of a single unity, usually coming from different data entry requirements, different data entry times, or a heterogeneous schemata (NAUMAN; HERSCHEL, 2022).

2.1.4 Data Quality Control

Validation, particularly relevant since Big Data became ubiquitous to the functioning of day-to-day activities, due to the fast velocity of arriving data and a large variety of heterogeneous sources, the quality of such data is far from perfect. Data quality assurance is the process of profiling data to discover inconsistencies, inaccuracies, incompleteness, and other data anomalies. In addition to performing data cleansing activities, data aggregation and data transfers can also be used to improve data quality (GAO; XIE; TAO, 2016). Basic data validation functions include: Null Data Value, Regex(Regular Expressions), Data type check, Data Range Validation, Data Constraint Validation (GAO; XIE; TAO, 2016).

2.2 Machine Learning

Machine learning is defined as an automated process that extracts patterns from data (KELLEHER; NAMEE; D'ARCY, 2020). A search through a set of possible prediction models for the model that best captures the relationship between the descriptive features and the target feature in a data set.

2.2.1 Types of Machine Learning

In supervised learning, the instances used during the training phase are labeled. The learning happens automatically from the relationship between a set of descriptive features and a target feature based on a set of historical examples, or instances.

Unsupervised machine learning techniques are used in the absence of a target feature and to model the underlying structure within the features in a dataset. Such inherent characteristics can make it specially suitable for clustering algorithms.

Semi-Supervised Learning is a branch of ML that aims to combine the previous two tasks (ENGELEN; HOOS, 2020). Typically, semi-supervised aims to improve results by gathering information normally associated with on task to improve the performance of the other.

Reinforcement learning is used mainly to learn to control the behavior of autonomous systems. It relies less on a dataset and more on the particular property of being able to recreate tasks in an environment repeatedly.

2.2.2 Key Concepts

A Feature is any measure derived from a domain concept that can be directly included in an Analytics Base Table (ABT) for use by a ML algorithm (KELLEHER; NAMEE; D'ARCY, 2020). Feature design is non-trivial, usually based on the availability, timing, and longevity of the data in question.

An ABT is usually made up of descriptive features created from operational data bases, files, external feeds, and the remaining fonts of data deemed relevant by the specialist in the field of application; and a characteristic of the target (KELLEHER; NAMEE; D'ARCY, 2020). The prediction subject defines the basic level at which the predictions

are made, having one per role and multiple descriptive features that are judged to be able to accurately predict their respective subject.

Data preparation can at times mean the removal of Noise or Artifact, but it also encompasses techniques that change the way data is represented with the aim of making it more compatible with certain machine learning algorithms. Binning and Normalization are some of the more common ones.

Continuous features in an ABT that cover different ranges can cause difficulty for some ML algorithms. Normalization techniques can keep the relative differences between the values while fitting them in the needed range.

$$a'_i = \frac{a_i - \min(a)}{\max(a) - \min(a)} \times (\text{high} - \text{low}) + \text{low}$$

where a'_i the normalized feature value, a_i is the original value, $min(a)$ is the minimum value of feature a, $max(a)$ is the maximum value of feature a, and low and high are the minimum and maximum values of the desired range (KELLEHER; NAMEE; D'ARCY, 2020). Typical ranges used to normalize feature values are [0,1] and [-1,1]. Binning contains converting a continuous feature into a categorical feature, where the range in which there are data is segmented, and each value instance has its value substituted by the corresponding category. The bin can be equal width, same delta of values for each bin, or equal frequency, same number of instances in each bin (KELLEHER; NAMEE; D'ARCY, 2020).

Model evaluation aims to answer if the model has a result above a "good enough" threshold while performing the function it was designed to perform. That process can be useful when comparing the suitability of different models for the same task, to estimate performance before deployment, or to convince third parties of employers that the model will be able to fulfill their needs (KELLEHER; NAMEE; D'ARCY, 2020). In these scenarios, the term Over and Under Fitting is often used, Over fitting occurs when an algorithm performs well on training but poorly on testing and under fitting when an algorithm has poor performance on both datasets (KELLEHER; NAMEE; D'ARCY, 2020).

Inductive bias is the set of assumptions that defines the model selection criteria of a ML algorithm. Inductive bias ML algorithms can be subdivided into restriction and preference bias. A restriction bias constrains the set of models that the algorithm will consider during the learning process. A preference bias guides the learning algorithm to prefer certain models over others (MINING, 2006).

2.2.3 Algorithms

The simplest algorithm that is still capable of expressing the ML mechanisms is the linear regression, it is a statistical model that, from the assumption that a linear relation between features exists, a weight (i.e., model parameter) is defined as such that when applied to a feature it is able to model another. As such, it is able to estimate missing sections of the data or predict future data.

The decision tree is another quintessential algorithm to ML, named after it's data structure: it consists of a root node, interior nodes, and leaf nodes, they are connected by branches. The number of possible levels that a descriptive feature can take determines the number of downward branches from a non-leaf node.

Support vector machines (SVMs): is a predictive model approach that is based on error-leaning. The objectives are to find a decision boundary that leads to the maximum margin that will best separate the levels of the target feature, measured by the distance of the closest instance of each cluster to the boundary (KELLEHER; NAMEE; D'ARCY, 2020).

2.3 Network Security

With the growing amount of data and the ubiquitous dependence of internet access, the perspective of network security has been a ever more important area of concern. A secure network is usually considered this when it is able to maintain a couple of proprieties: confidentiality, integrity of messages, endpoint authentication, and operational security. In a network that maintains confidentiality, only the sender and the receiver should be able to understand a message. Since a network encompasses the traffic of multiple possible senders and receives, thus allowing the interception of a message, to maintain confidentiality, some form of encryption is necessary. As for a network to maintain the property of message integrity, a message sent should be the same as the one received, and no alteration should happen during it's transit. Measures to maintain this propriety must account for possible malicious and accidental causes. For it to be authenticated at the endpoint, a network must allow the sender and the receiver to be able to confirm the identity of each other. Finally, for the principle of operational security to be respected, there must be secure control of the access between information sent and received from the Internet to the local network.

2.3.1 Threats and Vulnerabilities

Threats can usually be classified, as defined by (STALLINGS; BROWN, 2015), by which property a secure network should have, compromised. A threat that compromises confidentiality is called an Unauthorized Disclosure, and it can happen by way of: Exposure, either intentional or caused by human, machine, or software error, is when private data is accidentally exposed; Interception when an attack occurs that allows some device to receive packets intended for another device; Inference when extra information about a network can be deduced based on available information; and Intrusion when an attack is able to overcome a system's access control protections.

Deception is a threat to the integrity of the system or the integrity of the data. It can be caused by: Masquerade attacks when an unauthorized user poses as an authorized user; Falsification is when real data is substituted by false data; Repudiation when a user sends, receives, or possesses data.

Disruption is a threat to the system's availability or integrity. It might be caused by: Incapacitation attacks, an attack to the system availability; Corruption, an attack to the system's integrity; or obstruction, either overload or obstruction of the system resources with the goal of compromising communication links.

Usurpation is a threat to the systems integrity, it encompasses Misappropriation and Misuse, that means theft of service (such as Denial of Service Attack i.e., DDoS) and this malicious logic or unauthorized access disabling security functions.

2.3.2 Security controls

Network segmentation principles state that resources of different security levels and requirements should be places in different zones protected by firewall implementing appropriate policies. It can be intuitively explained by the rationale that different parts of a system possess different levels of risk and tolerances and, as such, should be placed in different security zones (MHASKAR; ALABBAD; KHEDRI, 2021).

Isolation in security has similar purpose to that of isolation in fault security, e.g. if a bug crashed one partition, it does not crash the whole system, that also applies in security but in addition to that, it is also expected that if one partition gets compromised, accessed by devices that should not have been accessed by, it should guarantee that the attack cannot be used to gain access to the other partitions (SHU et al., 2016).

AAA (Authentication, Authorization and Accounting) mechanisms are a framework that aims to, in a scalable manner, coordinate network technologies preserve security. Authentication involves validating end users' identity before allowing access to the network. Authorization defines what privileges, access levels and actions one has once it has access to the network. Accounting provides methodology for properly monitor each user use of the network, be it for charging, development planning, and so on (METZ, 1999).

3 METHODOLOGY

The literature review methodology aims to be an evidence-based research and is used to identify, evaluate, and summarize the available research relevant to a group of research questions (KITCHENHAM; CHARTERS et al., 2007). It was first adopted in medicine after research indicated that expert opinion-based medical advice was not as reliable as advice based on the accumulation of results of scientific experiments (KITCHENHAM et al., 2009). Since then, it has been adopted in different research domains: urban planning (COCCHIA, 2014), marketing (BOCCONCELLI et al., 2018) and project management (AARSETH et al., 2017).

The Systematic Literary Review (SLR) methodology used in this work as a reference is the tailored methodology for computer science research, defined in (CARRERA-RIVERA et al., 2022) and from which the formal definitions for this study were sourced. This review method comprises two main phases **Planning** and **Conducting**. Planning involves defining the protocol, starting with the PICO elements, and generating the search query based on them.**Conducting** involves following, executing and recording the steps defined in the first phase. These phases are described in the next sections.

3.1 Planning

The first step in an SLR is defining the protocol, since it describes procedures and can act as a log of the activities to be performed with the goal of allowing the reproducibility of the reviews. The review process consists of six steps:

1. Define PICO Elements
2. Formulate Research Questions
3. Select Digital Libraries
4. Define Inclusion and Exclusion Criteria
5. Define Quality Assessment
6. Define Extraction Form

3.1.1 PICO Elements and Synonyms

The PICO (Population, Intervention, Comparison, Outcome) breaks down the objectives into searchable keywords and synonyms that help formulate research questions (PETERSEN; VAKKALANKA; KUZNIARZ, 2015). Table 3.1 presents the definitions of the elements used in this study.

Table 3.1: Define PICO elements and synonyms

	Description	**Keyword**	**Synonyms**
Population	Can be a specific role, an application area, or an industry domain.	Data Quality, Machine Learning, Network Monitoring	AI, IoT, Distributed Systems, Network Security.
Intervention	The methodology, tool, or technology that addresses a specific issue.	DQ Applied to ML or Networks	ML DQ metrics, Network oriented DQ metrics
Comparison	The methodology, tool, or technology in which the Intervention is being compared (if appropriate).	Traditional DQ metrics	DQ Metrics
Outcome	Factors of importance to practitioners and/or the results that Intervention could produce.	Metrics more precise than the generic ones	DQ assessment

3.1.2 Research Questions

The research questions are the guiding light of the search and filtering steps that succeed. They are the basis for the search terms and the filtering criteria; as such, must accurately represent the knowledge that the review aims to bring forth. In Table 3.2 are presented the four Research Questions (RQ) that guided this review.

RQ1 and RQ2 were used as a way to represent one of the two main goals behind this research, establish how DQ has been used, if it is tailored to fields of application (and in which fields data quality assessment (DQA) is more prevalent. On the other hand, RQ3 and RQ4 represent a smaller scope and more focused guide, being very precise in trying to find only the extremely relevant information to the topic that prompted this literature review.

Table 3.2: Research Questions

#	**Research Question**
RQ1	Is there a consensus on the metrics to analyze DQ?
RQ2	In which field of application has DQ research been the most prevalent?
RQ3	How is DQ monitored or considered in ML research?
RQ3	Are the efforts to quantify DQ for ML in the network security context?

3.1.3 Digital Libraries Sources

This review will only be as representative as the reality of the current research as the universe it is trying to synthesize. Keeping that premise in mind, two of the largest multidisciplinary science literature repositories were chosen, Web of Science and Scopus. To supplement the breath of search, databases specifically relevant to computer science research were also added, in ACM Digital Library and IEEE Xplore. Table 3.3 summarizes the main aspects of each library.

Table 3.3: Digital Libraries Used in the Research

Database	**Description**	**Area**	**Advanced Search**
IEEE Xplore	Scientific and technical article repository	Technology	Yes
ACM Digital Library	Computing and information technology article database	Computing and Information Technology	Yes
Scopus	Multidisciplinary article and abstract database	Interdisciplinary	Yes
Web of Science	Multidisciplinary article and abstract database	Interdisciplinary	Yes

3.1.4 Inclusion and Exclusion Criteria

The inclusion criteria were defined before starting the review, as shown in Table 3.4, to prevent bias, although some adjustments were made during the process. The first round of exclusion was based on the abstract and primary bibliographic data. In case the decision was unclear, the article was skimmed to lead to a more informed decision.

Table 3.4: Inclusion and Exclusion Criteria

Type	*Criteria*
Inclusion	Abstract indicates that the work is: • Strongly related: DQ assessment or management, aiming to improve or predict ML training **or** Network health monitoring and improvement. • Weakly related: DQ assessment tailored for a specific field of application.
Exclusion	• Repeated from another database. • Topics of interest are only tangentially mentioned. • Year of publication is prior to 2015.

3.1.5 Quality Assessment Checklist

The second round of exclusion, Quality Assessment (QA), used a checklist that includes three evaluation points. A numerical scale was used to assess the criteria and quantify the QA, the same scale was used for all established criteria. The following table 3.5 further specifies how the filtering process evaluation was done. Though the broad scope of this methodology was defined during the planning phase, some modifications were done during the conducting phase to assure that the cutoff was being truly representative of the research's initial intent. After the QA process, six publications were selected by the defined criteria, to represent the literature and provide relevant answers to the defined research questions.

3.1.6 Data Extraction Form

Table 3.6 defines the desired metadata from the studies that remain after the filtering step based on abstract and keywords.

3.2 Conducting

The following steps comprise the execution of the procedures defined in the protocol. Starting by the definition of the search string queries, the gathering of the publica-

Table 3.5: QA Assessment Checklist

Type	*Criteria*
Questions	Level of participation: 1. Does this article use DQ applied to ML or Network Security? 2. Does this article define the DQ metrics used in the experiment application? 3. Does this article explain or defend the reasons or motivations behind the chosen data quality assessment protocol?
Answers	To assign a weight for each question answered. The weight is: • 1.0 if the answer is "YES", • 0.5 if the answer is "Partially", • 0.0 if the answer is "NO".
Cutoff Score	The cutoff is 70% of the maximum possible score. The score is the added values of the answers for the questions for each publication. Papers whose total score is less than the cutoff are not considered for the deep review phase.

tions, and the selection and refinement.

The first papers, surveys, and articles (that passed the process of review of the abstracts). This first process resulted in 80 articles that were then further filtered by applying the exclusion criteria in Table 3.4. Finally, the scoring method defined in Table 3.5 was used as the last selection process that resulted in the final 6 articles, the cutoff score choice was made in accordance with (CARRERA-RIVERA et al., 2022).

Table 3.6: Data Extraction Form

#	**Data Extraction Form**
Title	Title as is, as published
Authors	Complete list of authors
Year	Year it was published
Event	Conference, journal, or magazine name
Event's Qualis	Grade conferred by CAPES, if any
Type of Paper	Article, Tutorial, Survey, Short paper
Field of Application	Main area of effect
Access	Open access or subscription required
Database	Database via which data was found
Link URL	URL to the data

3.2.1 Digital Library Search Strings

The search strings were built considering the PICO elements. Resulting the following search strings:

1. `"data quality"`
2. `"data quality AND (machine learning OR artificial intelligence)"`
3. `"data quality AND (computer networks OR networking OR distributed systems OR computer security OR IoT OR Internet-of-Things)"`

3.2.2 Gathering Publications

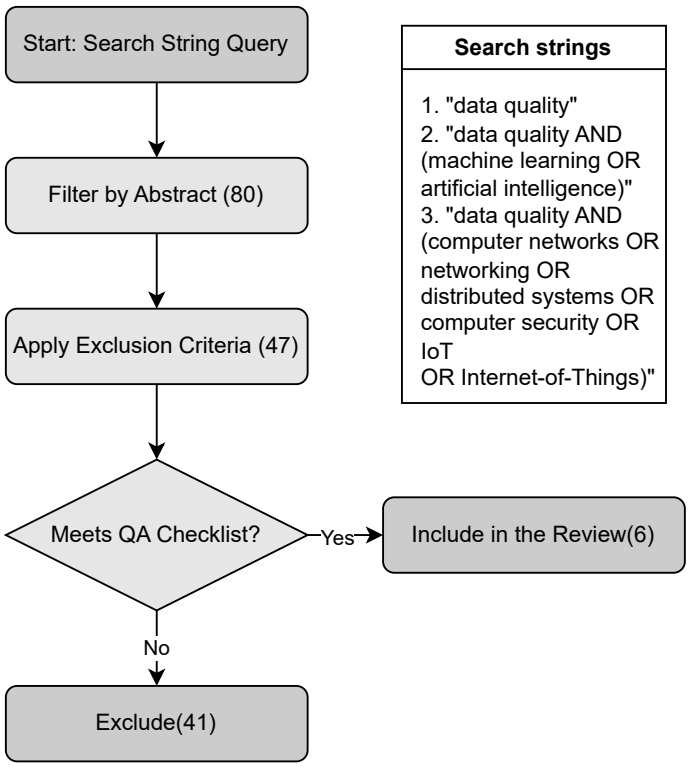
These queries were performed in the digital libraries resulting in a total of 80 publications which the abstracts related to the main topics. Then, each one of the publications was filtered using the criteria defined in Section 3.1.4.

3.2.3 Study Selection and Refinement

After the first contact with the data, when the abstract was decided and in some cases a skimmed read of the article if it would be relevant for this review, some metadata about the selected portion of the publications was collected, shown in Table 3.6 are which

information about the publications[1] was archived and how it was classified, in a way that helped the next step of Applying the exclusion and inclusion criteria. After conducting these steps, 6 publications were left that were analyzed in-depth, which are described in the next chapter. The steps in a graphical form are depicted in Figure 3.1.

Figure 3.1: Literature Review Flow Followed in this Work



[1]The spreadsheet with the selection process documentation can be found in https://github.com/lucaskruger/Data-Quality-in-Artificial-Intelligence-and-Computer-Networks-A-Systematic-Literature-Review.git

4 REVIEWS OF SELECTED PUBLICATIONS

In this chapter, each of the publications selected from the filtering process established in the previous chapter is presented individually, in a concise manner, with a focus on the goal and initial assumptions and the conclusion of the experiment. Such priorities can, at times, lead to the absence of methodology details or steps of the experiments, for the sake of succinctness and clarity. An exception will only be made when process itself is directly related to the RQ3.

4.1 The Effects of Data Quality on Machine Learning

Starting from the known dependence that modern Artificial Intelligence (AI) has on large quantities of available and reliable data, the study empirically explores the relation of 6 DQ metrics on the training data, test data, or both) on the performance of 19 Machine Learning (ML) algorithms, encompassing classification, regression, and clustering. The metrics used are Consistency, Completeness, Feature Accuracy, Target Accuracy, Uniqueness and Class Balance (BUDACH et al., 2022).

For the classification algorithms, uniqueness, consistent representation, and target class balance (if the class balance is not shifted to the extreme) have the lowest impact among the metrics tested. Suggesting that for a low loss in classifier performance on tabular data, skipping preprocessing steps like exact deduplication, unifying inconsistent representation, and carefully balancing the target variable may be a reasonable choice. For regression algorithms completeness, feature accuracy and target accuracy are the ones with the biggest impact, with a worse than linear decrease in algorithm performance. Missing values or inaccurate features in the test data without having trained the model on such type of data leads to even worse performance. Uniqueness and Target Class balance show little impact. Consistent representation only significantly impacts when the new representation outweighs the old ones. For Clustering, Completeness and Feature Accuracy seems to be with the most ones, leading to at least linear degradation of clustering performance. Surprisingly, the other four metrics had very little impact.

4.2 Data Curation and Quality Evaluation for Machine Learning-Based Cyber Intrusion Detection

Starting from the lack of consideration for data quality in other models aimed at detecting intrusions from large amounts of data, this study investigates how DQ may affect ML performance, through experiments on 11 host-based intrusion datasets, using eight ML models and two pre-trained language models. The language models, BERT and GPT-2 outperformed, being less susceptible to duplicates and overlaps than the traditional ML models. The second test showed that evaluating the pre-trained models and ML models on duplicate data does lead to unreliable evaluations. Furthermore, removing overlaps within a range of sequences could improve the performance of pre-trained models, at the cost of possibly reducing performance in datasets with highly similar sequences (TRAN et al., 2022).

4.3 Data Quality for Software Vulnerability Datasets

Automatic software vulnerability automatic detection relies on large software vulnerability datasets, as much for testing as for benchmarking, thus the potential negative impact of data quality in such datasets is unknown. This study, inspecting five inherent DQ attributes for four software vulnerability datasets, found that 20 to 71% of the vulnerability labels were inaccurate to the real world datasets and 17 to 99% of the data points were duplicated, those having significant impact on downstream models, either compromising training or testing (CROFT; BABAR; KHOLOOSI, 2023).

The metrics used were Accuracy, Uniqueness, Consistency, Completeness and Currentness. Issues in Uniqueness, Consistency and Completeness could be detected by rule-based syntactic filters. Since accuracy filtering could potentially decrease the number of entries in the data by 20-71%, rendering it too small for the use in question. Since no sufficiently good treatment to improve DQ was found in the study, the faults of in collection identified were that: Automatic data collection often leads to data inaccuracy, source code duplication may make datasets lack diversity,and unknown vulnerabilities can introduce label inconsistency.

4.4 Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations

Starting with the statement that most of the traditional DQ literature encompasses relation data, this work aims to provide the main challenges and differences when dealing with the concept and application of DQ in Big Data and ML. Big data especially deals with many instances of missing data, if there are too many in a critical attribute, deletion is an option that will have an effect on the statistical power (GUDIVADA; APON; DING, 2017).

Though big data DQ issues are presented, the ML section mainly covers ML general concepts without the desired focus on fields specific considerations this review is most interested in.

4.5 Data Quality Based Intelligent Instrument Selection with Security Integration

Proposes the idea of Data Quality with Security (DQS). Develops a framework for integration of heterogeneous data from multiple sources, aggregates securities metrics into the integral DQS calculus, particularly focused on the scenario of a network of sensors. DQ evaluation is performed mainly via intrinsic metrics. The DQS is flexible to account for the differences in the possible sensor networks to which it can be applied, some of the metrics used are: Sensor Accuracy(SA) related to precision and in the case study presented refers to the sensor's ability to capture details or clarity in the received signal, and the total sensor accuracy that is based on the SA on all the sensors in the platform (CHUPROV et al., 2024).

Sensor Latency is the average o minimum and maximum delays the sensor shows between its measurements. Instrumental Power Consumption is based on the power consumed when measuring and when idling. Now as for the security metrics Instrumental Platform Security is used, since the test case employs Android OS mobile devices, the metrics used were: "screen lock activation status" and "Android OS basic integrity test" among others.

4.6 Unsupervised Anomaly Detection in Data Quality Control

Starting from the statement that organizations typically detect data errors manually and reactively, making is time consuming and and prone to human error, this article proposed a automated anomaly detection approach, based on models comparison, consensus learning, and a combination of rules of thumb. It makes use of accuracy, completeness, consistency, and timeliness to access data quality. Though it concludes that fully automatic data quality control is not yet feasible, as even their proposed model still requires a human with specific knowledge of the field of application to help judge the evaluation process, it backs the validity of consensus for unsupervised learning by showing a precision score of 0.922 using their ensemble learning technique, a promising result (POON et al., 2021).

4.7 The Challenges of Data Quality and Data Quality Assessment in the Big Data Era

The data quality assessment protocol consists of five dimensions of DQ availability, usability, reliability, relevance, and presentation quality. The first four quality dimensions are considered indispensable and inherent features of data quality, and the final dimension is additional properties that improve customer satisfaction (CAI; ZHU, 2015).

The complete process establishes a sequence of incremental definitions. A most relevant step is the usage of indicators; specific assessment indicators must be chosen for every dimension, Reliability for one can be based on Accuracy, Consistency, Integrity, and Completeness. Those choices must be made following the most relevant metric for the area of application. The following step defines the requirements that the data needs to comply to a baseline score needed in the dimensions evaluation previously defined. Only then should the data collection step proceed.

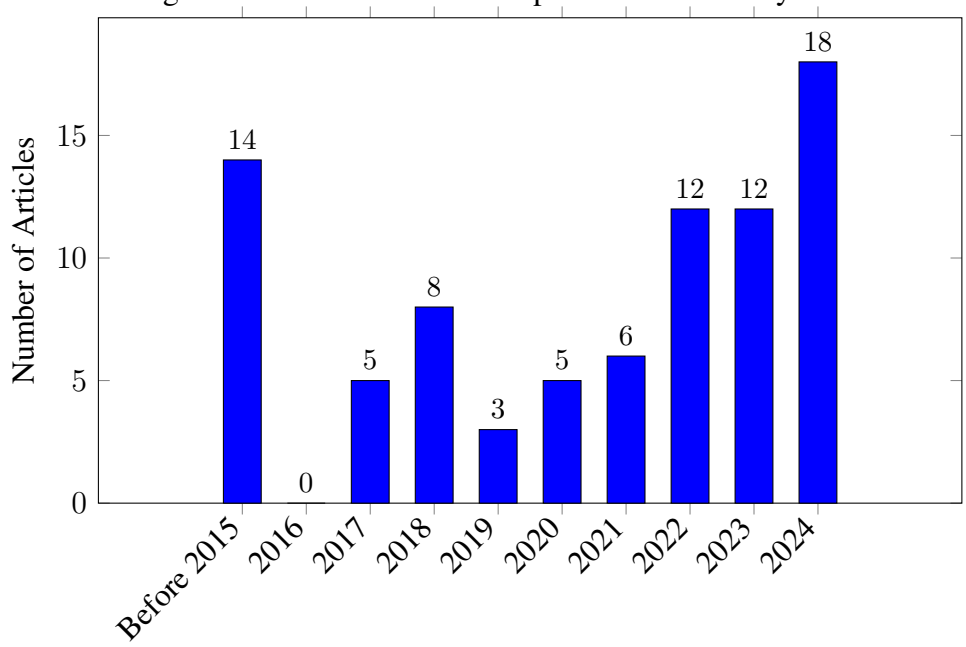
5 RESULTS AND RESEARCH QUESTIONS

This chapter presents the main results of the systematically oriented review of the literature and aims to answers the RQs posed for this work.

5.1 Publications by Year

The graph 5.1 represents the population of approaches after the first selection done via the publications' abstracts, which had then been selected if they indicated that the correspondent study could be relevant to the topic of interest.

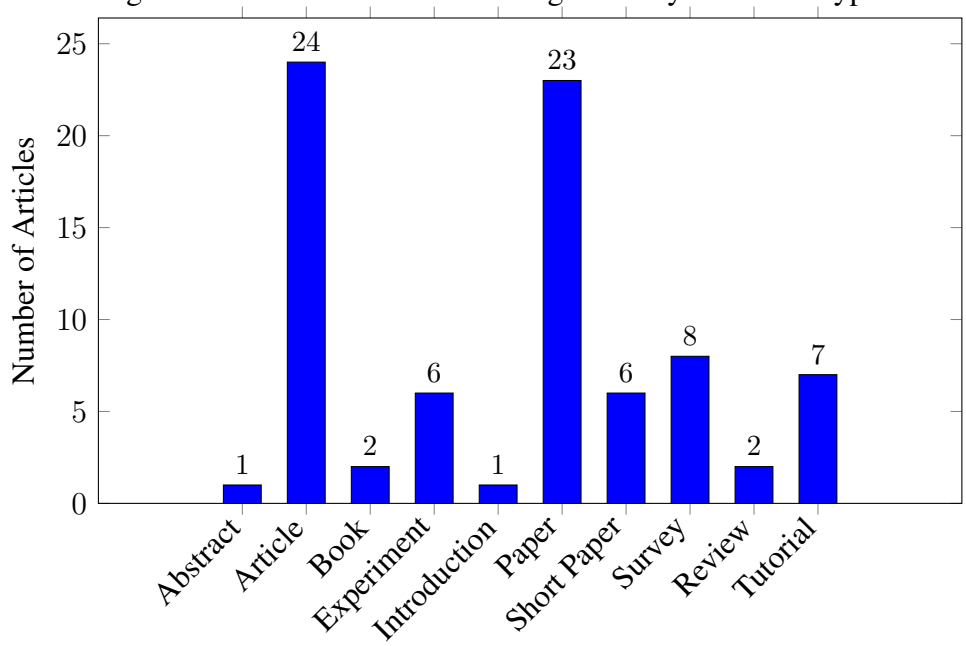
Figure 5.1: Number of articles published over the years



5.2 Publications by Type

The graph 5.2 represents, as in the last section, the population of works evaluated in the steps after the first selection; the terms were assigned based on how they were described in their respective databases.

Figure 5.2: Number of articles categorized by document type



5.3 Answering RQs

5.3.1 RQ1 - Is there a consensus on the metrics to analyze DQ?

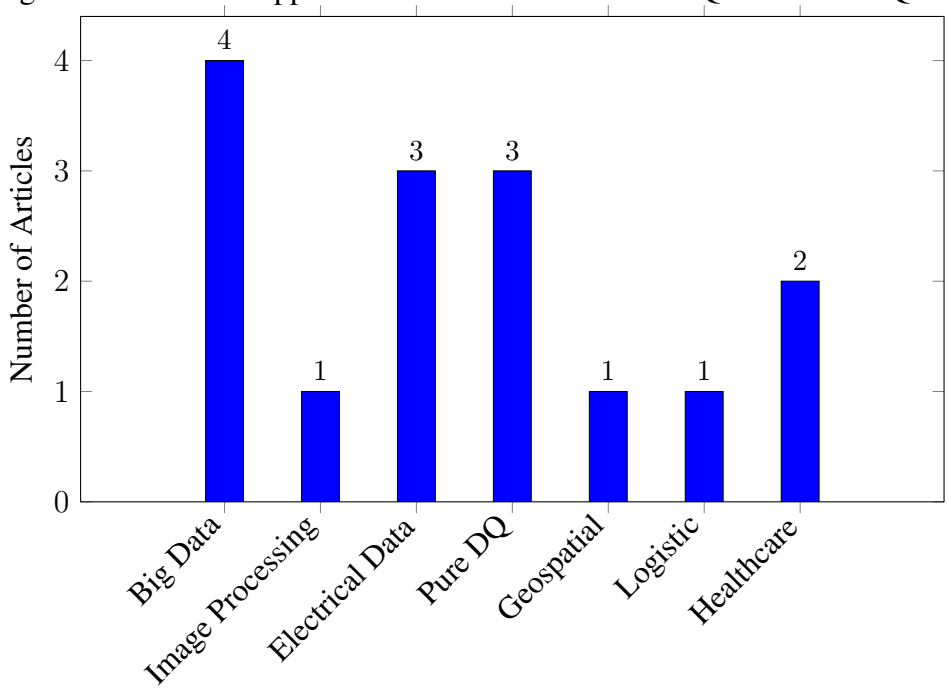
Although there is no set of metrics that can be observed in all the reviewed publications, a subset of intrinsic metrics does have a greater presence in completeness and accuracy, specifically having a high presence in the analyze publications, though usually complemented by other metrics, such as reputation and variety in (TRAN et al., 2022), or adapted, as in (BUDACH et al., 2022), to the field of application. Following the pattern (CHUPROV et al., 2024) uses accuracy, completeness, consistency, and timeliness and weights the input a specialist , of the filed of application, into the calculation.

Also, beyond the direct effect that DQ has in ML, publications that suggest Big Data handling techniques can also have an important indirect effect on ML algorithms performance. Many of the novel areas of application that ML is now very prevalent only became feasible when the available data started to be comprehensive enough to both allow the model to produce useful insights and require novel data handling techniques capable of dealing with its scope (ZHOU et al., 2017). As such, Big Data data quality assessment protocols may be relevant to the goals of this review. As such (CAI; ZHU, 2015) was manually included in the selected publications as a way to complement the scope of the reviews, but not deviate from the objectives of this study.

5.3.2 RQ2 - In which fields is DQ assessment the most prevalent?

Despite the focused nature of this review in the application, usage, of DQ, the overall state of DQ assessment process was included in a sub query, among the selected articles in first filtering stage the areas in which DQ is most relevant are shown in Figure 5.3. 65 publications in the network or ML area were found, but were not included in the figure since the search query that resulted in them already specified the field of application.

Figure 5.3: Field of Application of the Results of the DQ Assessment Query



5.3.3 RQ3 - How is DQ monitored or considered in ML research?

In (BUDACH et al., 2022) 6 DQ metrics were taken into consideration for testing, 5 of them being direct adaptations for ML of the traditional intrinsic metrics, though Accuracy was used as in two metrics, Target and Feature Accuracy. The sixth metric being Class Balance, based on the knowledge that clustering algorithms tend to miss less populated classes or known phenomenons from the algorithms themselves, e.g., k-means tending to generate uniform sized group even when that is not represented on the data.

In (TRAN et al., 2022) the main evaluation criteria are the following. Reputation means trustworthiness, calculated by statistical distance or by score attributed to datasets

based on referrals or reviews done by the community. Relevance represents if the data is indeed from a source representative of the area of interest. Comprehensiveness means if the data is representative of the whole population, especially relevant for ML and DL once a DL model contains millions of parameters and requires a large amount of data for training. Other contributing measurements used in the study are variety, timeliness, accuracy, consistency, duplication, and overlap.

Overall, these publications show the significant work put towards adapting DQ ideals from it's traditional definitions to ML, and at times specifically DL, data requirements.

5.3.4 RQ4 - Are there efforts to quantify DQ for ML in the network security context?

Yes, in (TRAN et al., 2022) it was attributed a high importance to data-quality control and it's consequences in ML models performance, as much as detailed analyses aiming at identifying useful correlations between the DQ metrics used and how much they could influence each of the ML algorithms, they also propose a high DQ assurance system specifically projected and tested for building a high-quality machine learning-based intrusion detection system.

Also, (CROFT; BABAR; KHOLOOSI, 2023) tested the impact that the DQ of vulnerability datasets had on learning-based techniques for detecting software vulnerability, and although its conclusions cannot be completely expanded to the context of network security, it does show results that may be relevant to future research. The consideration of improving overall DQ of a dataset at the cost of size, through deletion of other more elaborate techniques, had prohibitive consequences once learning models requires vast amounts of data, bringing the need for it's collection process to be automated and that for it's turn leads to more inaccurate data. In general, the process in the methodology (CROFT; BABAR; KHOLOOSI, 2023) and the elucidation of the steps taken can prove a valuable resource for publications on DQ for ML in a network security context.

6 CONCLUSION AND FUTURE WORK

This study set out to explore the current literature that could possibly help in the development of a ML-based tool, capable of monitoring network height in the data quality of the data that it is monitoring. Also, a related interest in the ways that DQ has been used in different areas of application has also been prioritized. Through the selection process, guided by an established methodology that aimed at reducing bias while also allowing some flexibility for scope resizing during the process, the articles that could best answer the research questions were selected.

Most of the DQA methodology found was a direct adaptation of the traditional metrics, common in relational databases evaluations, that were complemented by a specific metric made based literature of that field. This scenario causes that those directly adapted to have a bigger weight in the overall quality evaluation.

The findings of this review aim to contribute to a deeper understanding of the DQA methodology and provide starting point options for the desired purpose, adapted from the reviews publications. Future works include proposing and testing new metrics to quantify the DQ of datasets in different fields, such as network security, network management, and network optimization; as well as involving explicability concepts into the metrics evaluation process.

REFERENCES

AARSETH, W. et al. Project sustainability strategies: A systematic literature review. **International journal of project management**, Elsevier, v. 35, n. 6, p. 1071–1083, 2017.

BOCCONCELLI, R. et al. Smes and marketing: a systematic literature review. **International Journal of Management Reviews**, Wiley Online Library, v. 20, n. 2, p. 227–254, 2018.

BUDACH, L. et al. The effects of data quality on machine learning performance. **arXiv preprint arXiv:2207.14529**, 2022.

CAI, L.; ZHU, Y. The challenges of data quality and data quality assessment in the big data era. **Data science journal**, v. 14, p. 2–2, 2015.

CARRERA-RIVERA, A. et al. How-to conduct a systematic literature review: A quick guide for computer science research. **MethodsX**, Elsevier, v. 9, p. 101895, 2022.

CHUPROV, S. et al. Data quality based intelligent instrument selection with security integration. **ACM Journal of Data and Information Quality**, ACM New York, NY, v. 16, n. 3, p. 1–24, 2024.

COCCHIA, A. Smart and digital city: A systematic literature review. **Smart city: How to create public and economic value with high technology in urban space**, Springer, p. 13–43, 2014.

CROFT, R.; BABAR, M. A.; KHOLOOSI, M. M. Data quality for software vulnerability datasets. In: IEEE. **2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)**. [S.l.], 2023. p. 121–133.

EMMANUEL, T. et al. A survey on missing data in machine learning. **Journal of Big data**, Springer, v. 8, p. 1–37, 2021.

ENGELEN, J. E. V.; HOOS, H. H. A survey on semi-supervised learning. **Machine learning**, Springer, v. 109, n. 2, p. 373–440, 2020.

FAN, W.; GEERTS, F. **Foundations of data quality management**. [S.l.]: Springer Nature, 2022.

GAO, J.; XIE, C.; TAO, C. Big data validation and quality assurance – issuses, challenges, and needs. In: **2016 IEEE Symposium on Service-Oriented System Engineering (SOSE)**. [S.l.: s.n.], 2016. p. 433–441.

GUDIVADA, V.; APON, A.; DING, J. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. **International Journal on Advances in Software**, v. 10, n. 1, p. 1–20, 2017.

HAEGEMANS, T.; SNOECK, M.; LEMAHIEU, W. Towards a precise definition of data accuracy and a justification for its measure. In: . [S.l.: s.n.], 2016.

HERNÁNDEZ, M. A.; STOLFO, S. J. Real-world data is dirty: Data cleansing and the merge/purge problem. **Data mining and knowledge discovery**, Springer, v. 2, p. 9–37, 1998.

KELLEHER, J. D.; NAMEE, B. M.; D'ARCY, A. **Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies**. [S.l.]: MIT press, 2020.

KIM, W. et al. A taxonomy of dirty data. **Data mining and knowledge discovery**, Springer, v. 7, p. 81–99, 2003.

KITCHENHAM, B. et al. Systematic literature reviews in software engineering–a systematic literature review. **Information and software technology**, Elsevier, v. 51, n. 1, p. 7–15, 2009.

KITCHENHAM, B.; CHARTERS, S. et al. **Guidelines for performing systematic literature reviews in software engineering**. [S.l.]: UK, 2007.

LI, C. et al. Noise filtering to improve data and model quality for crowdsourcing. **Knowledge-Based Systems**, Elsevier, v. 107, p. 96–103, 2016.

METZ, C. Aaa protocols: authentication, authorization, and accounting for the internet. **IEEE Internet Computing**, IEEE, v. 3, n. 6, p. 75–79, 1999.

MHASKAR, N.; ALABBAD, M.; KHEDRI, R. A formal approach to network segmentation. **Computers & Security**, Elsevier, v. 103, p. 102162, 2021.

MINING, W. I. D. **Introduction to data mining**. [S.l.]: Springer, 2006.

NAUMAN, F.; HERSCHEL, M. **An introduction to duplicate detection**. [S.l.]: Springer Nature, 2022.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. **Information and software technology**, Elsevier, v. 64, p. 1–18, 2015.

PIPINO, L. L.; LEE, Y. W.; WANG, R. Y. Data quality assessment. **Communications of the ACM**, ACM New York, NY, USA, v. 45, n. 4, p. 211–218, 2002.

POON, L. et al. Unsupervised anomaly detection in data quality control. In: IEEE. **2021 IEEE International Conference on Big Data (Big Data)**. [S.l.], 2021. p. 2327–2336.

SCANNAPIECO, M. **Data quality: concepts, methodologies and techniques. Data-centric systems and applications**. [S.l.]: Springer, 2006.

SHU, R. et al. A study of security isolation techniques. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 49, n. 3, p. 1–37, 2016.

STALLINGS, W.; BROWN, L. **Computer security: principles and practice**. [S.l.]: Pearson, 2015.

TRAN, N. et al. Data curation and quality evaluation for machine learning-based cyber intrusion detection. **IEEE Access**, IEEE, v. 10, p. 121900–121923, 2022.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: What data quality means to data consumers. **Journal of management information systems**, Taylor & Francis, v. 12, n. 4, p. 5–33, 1996.

XIONG, H. et al. Enhancing data analysis with noise removal. **IEEE Transactions on Knowledge and Data Engineering**, v. 18, n. 3, p. 304–319, 2006.

XU, D.; ZHANG, Z.; SHI, J. A data quality assessment and control method in multiple products manufacturing process. In: **2022 5th International Conference on Data Science and Information Technology (DSIT)**. [S.l.: s.n.], 2022. p. 1–5.

ZHOU, L. et al. Machine learning on big data: Opportunities and challenges. **Neurocomputing**, Elsevier, v. 237, p. 350–361, 2017.