Received: 24 June 2020 | Revised: 28 March 2021 | Accepted: 30 April 2021

DOI: 10.1002/cpe.6385

RESEARCH ARTICLE

WILEY

A feature-based approach for guiding the selection of Internet of Things cybersecurity standards using text mining

Koen van der Schaaf[1] | Bedir Tekinerdogan[1] | Cagatay Catal[2]

[1]Information Technology Group, Wageningen University and Research, Wageningen, the Netherlands

[2]Department of Computer Science and Engineering, Qatar University, Doha, Qatar

Correspondence
Cagatay Catal, Department of Computer Science and Engineering, Qatar University, Doha, Qatar.
Email: ccatal@qu.edu.qa

Abstract

Cybersecurity is critical in realizing Internet of Things (IoT) applications and many different standards have been introduced specifically for this purpose. However, selecting relevant standards is not trivial and requires a broad understanding of cybersecurity and knowledge about the available standards. In this study, we present a systematic approach that guides IoT system developers in selecting relevant cybersecurity standards for their IoT projects. The systematic approach has been developed in four stages. First, the common and variant features of IoT cybersecurity have been modeled using a feature model. Second, an up-to-date overview of the IoT cybersecurity standards landscape has been mapped by combining existing overviews. Third, a text mining algorithm has been implemented. Fourth, the systematic approach has been modeled using business process modeling notation. Our case study demonstrated that this approach is effective and efficient for guiding the selection of IoT cybersecurity standards.

KEYWORDS

cybersecurity, feature model, Internet of Things, natural language processing, standards, text mining

1 | INTRODUCTION

The *Internet of Things* (IoT) is defined as a network of physical objects or things equipped with hardware, software, sensors, and network connectivity components. Through the usage of these components, the physical objects can collect and exchange data with each other and humans over the Internet.[1,2] The possible applications where the IoT can be utilized span across multiple domains such as health, manufacturing, agriculture, transportation, and home automation.[2–4] Due to the many applications IoT devices can fulfill, they are becoming embedded in our society. However, these ever-present and always-connected devices cause major security threats when they are not secured.[5]

Currently, the security level of many IoT devices available on the market is not up to standard, and many devices can easily be compromised by attackers.[5,6] This poses several threats, including the leakage of sensitive data, devices that are controlled by intruders, and the creation of networks of compromised devices (i.e., botnets).[7] IoT botnets can be used for massive distributed denial of service attacks. This occurred in 2016, when an IoT botnet attacked Domain Name Server provider, DynDNS, which resulted in major websites such as Twitter, The Guardian, Netflix, and CNN being unavailable.[8]

For IoT applications to be widely adopted and accepted by the public, these security and privacy issues have to be adequately addressed.[9] Traditionally technology standards have played an essential role in the uptake of technologies,[10] and for the IoT, this is not different.[11,12] Standards organizations (SO) such as ETSI, IETF, and ISO are already publishing and developing cybersecurity standards specifically aimed at the IoT or technologies used in IoT applications. However, both researchers and industrial experts have noted that the amount and diversity of standards are too

complex to oversee. For example, Reference 5 states that there is no consensus among the standard-setting community,[11] notes that global collaboration between standard-setting organizations is required to solve the lack of consistency among the organizations and their standards, and in Reference 13 several industry leaders state that there is little consistency among the standards landscape. This corresponds with the findings from Reference 14, where firms indicated that the complexity of the IoT cybersecurity landscape made it difficult to select appropriate standards for their IoT system. The objective of this study is to develop a systematic approach for guiding the selection of cybersecurity standards. Based on this objective, this article provides answers to the following research questions.

- RQ1. *What are the common and variant features of IoT cybersecurity?*
- RQ2. *Which standards for IoT cybersecurity exist?*
- RQ3. *What is a feasible approach for automatically classifying IoT cybersecurity standards as one or multiple concepts from the domain analysis?*
- RQ4. *How can firms select appropriate IoT cybersecurity standards from a set of standards?*

This article contributes knowledge and insights on how firms can select appropriate IoT cybersecurity standards for their IoT system. The contributions of this article are the following:

- A systematic approach is presented that helps firms in selecting appropriate IoT standards for their system.
- A feature-oriented domain analysis (FODA) is applied to give insights into the common and variable features of IoT cybersecurity.
- The feature model allows developers of IoT systems to select the features they have or wish in their IoT system.
- IoT cybersecurity standards were combined in an aggregated overview and metadata on the standards was collected.
- To secure that future researchers or firms use the feature model and algorithm in a systematic way business process modeling notation (BPMN) diagrams were constructed of the FODA, and of a proposed business process for firms who are interested in finding relevant standards for securing their IoT system.
- A new classification algorithm is proposed and evaluated for textual data.

The remainder of the article is organized as follows: Section 2 presents the background, Section 3 introduces a feature-driven characterization of the IoT cybersecurity domain, Section 4 presents an approach for automatically classifying IoT cybersecurity standards, Section 5 presents the threats to validity, and Section 6 concludes this article.

2 | BACKGROUND AND RELATED WORK

2.1 | Cybersecurity

Cybersecurity has been defined differently throughout the literature. Schatz et al.[15] proposed an overarching definition; Lexical overlap and semantic similarity analyses were used to create a single definition that captures the essence of all the definitions in the dataset. This resulted in the following definition of cybersecurity, which is also the definition used throughout this article: *"The approach and actions associated with security risk management processes followed by organizations and states to protect confidentiality, integrity, and availability of data and assets used in cyber space."*.

Cybersecurity for both traditional IT systems and IoT systems has three main objectives: confidentiality, integrity, and availability, this is referred to as the CIA triad. The priority of these three objectives might be different between systems, as some systems will prioritize high confidentiality and others might prioritize availability. This will depend on the needs and wants of the stakeholders involved in the development of the IoT system. They are defined as follows[9]:

- Confidentiality refers to ensuring that the used and transferred data is only available to authorized users and cannot be monitored or interfered with by nonauthorized users.
- Integrity refers to ensuring protection against data modification and interference (or even destruction) and includes ensuring that data is authentic.
- Availability refers to ensuring that data and devices are available to authorized users and services when they are requested.

The characteristics of IoT systems make them more difficult to secure than traditional IT systems. Systems have to deal with potentially thousands of devices, which can autonomically interact with each other or other systems, these interactions have to be secured.[5] Securing IoT systems

can prove difficult because the devices are mobile, and the protocols, platforms, and devices are highly heterogeneous.[12] In addition, the systems might consist of devices that were not originally designed to be connected to the Internet, and parts of the system might be controlled by third-parties.[12] IoT devices are often constrained on resources, that is, they often have little amounts of processing power and energy. This results in difficulties in applying traditional security mechanisms such as encryption methods. For IoT devices, lightweight protocols have to be implemented, which has proven difficult for system developers. In addition, the devices can be deployed in an environment where they are unsupervised and exposed to potential physical tampering.

2.2 | IoT cybersecurity standards

Standards are developed by SO; these include standard-setting organizations such as ISO.[16] Standards in the technology industry are defined as a set of rules which describe the characteristics of a technology.[10,16] The definition of standards that are used in this article is based upon the work of Hogan and Piccarreta,[12] who researched the status of international cybersecurity standardization for the IoT.

A standard is defined as "*a document, established by consensus and approved by a recognized body, that provides for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context.*" [17, p. 43].

The scope of cybersecurity standards that are reviewed in this article is aligned with the study of Brass et al.[14] and is defined broadly to include principles, guidelines, codes of practice, and technical specifications. These can be developed by public, private, and not-for-profit organizations, which might include governmental institutions, (inter)national organizations, industry alliances, and associations. Over the past few years, several studies have provided insights on IoT cybersecurity standardization. To guide researchers, Riahi Sfar et al.[18] have identified major actors that participated in IoT security standardization and mapped their relevant activities. Keoh et al.[19] described the standardization efforts of the IETF for securing the IoT and concluded that these efforts are essential to make the usage of IoT applications a reality. Hogan and Piccarreta,[14] Keoh et al.,[19] and Lee et al.[20] aimed to map cybersecurity standards relevant for IoT systems to different areas of cybersecurity, review the main trends in the development and evolution of IoT cybersecurity standards, and provide an overview of the IoT standards landscape, respectively. These studies provided overviews of the IoT cybersecurity standards, which are unified in this article for further analysis.

2.3 | Natural language processing

This article uses natural language processing (NLP) techniques to retrieve the importance of keywords from a set of standards to classify each standard according to features derived from the feature model. To the authors' knowledge, no prior research classified cybersecurity standards using NLP. Preprocessing, stemming, feature selection, and term weighting, as described in Reference 21 was applied to the raw texts extracted from the standards. By applying preprocessing, stemming, and term weighting, the unstructured texts are transformed into structured data. The term weights calculated for the features indicate how important the feature is for the document and are used to classify the standards. Below preprocessing, stemming and term weighting for classification purposes are explained.

Preprocessing in NLP refers to the removal of noise, such as words and characters that add little or no information, from the texts. This is done to improve the performance and speed of the process.

Stemming is the automatic reduction of words to their stem, for example, the words generate, generated, and generates are stemmed to "generat." The Snowball stemmer from NLTK was used, this is an improved version of the popular Porter stemmer which works better for English. The stemming had to be applied to both the feature-queries and the texts from the PDF files. Unfortunately, stemming also creates some issues due to its rigid nature, which could result in invalid results.

It was decided to use features from the feature model as a search queries and count the number of times it appears in the text (i.e., determine its term frequency), these are referred to as "feature-queries." Extracting and processing and classifying the text using term frequencies is similar to a Bag of Words representation of a document. The context, location, and sequence of words are disregarded, and only the number of times a word appears in the document is of importance. For example, the document "*this document is about a standard*" and the document "*this standard is about a document*" are seen as identical. However, it is assumed that this will have little impact on the results of the classification, as a document covering a certain feature is more likely to mention the feature more often, thus resulting in a higher term frequency. Because raw term frequencies are unfair to use, as some words naturally appear more than others, they were weighted using a weighting scheme called term frequency-inverse document frequency (TF-IDF). For example, the terms from the CIA triad are often referred to when covering cybersecurity, naturally, these words will occur more than others. The TF-IDF score increases with the number of times a word occurs in a document combined with the rarity of the term. TF-IDF can be used for numerous analyses, such as document comparison and keyword extraction, in this article it was used for the classification of the standards.

This type of classification problem is considered a multilabel classification problem, also known as any-of classification. That is, a document can be classified to belong to several classes simultaneously, to a single class, or to none of the classes.[22] For example, it is possible that a standard covers how to implement multiple features in an IoT system, such as a certain network protocol and a certain access control mechanism, which indicates that it should be classified as covering both. It could also be possible that the standard only covers how to implement a certain network protocol, which should result in only a single classification.

A common method for text classification is to use machine learning classifiers, such as Naive Bayes or K-nearest neighbors. However, to develop these classifiers for text documents requires large amounts of training data, documents must be manually assigned to classes, which requires a lot of manual effort to do properly. This is unpractical, and in the context of this research also undesirable. The field of IoT cybersecurity is still evolving and is not fully established yet. The probability that new features can be added to the feature model in the near future is relatively high, which would require manual effort for categorizing new standards, add these to the training data, and retraining the model. This problem is known as the Knowledge Acquisition Bottleneck. Therefore, the standards were classified on the term frequencies of the feature-queries. This is a flexible solution, where new features can be added, or old features can be removed easily.

3 | FEATURE DRIVEN CHARACTERIZATION

In order to identify the possible classifications that cybersecurity standards could fall into, a domain analysis was conducted. Domain analysis is defined as identifying, collecting, organizing, and representing the relevant information in a domain.[23] Based on domain analysis, 15 influential papers were analyzed, and the common and variable concepts of each paper were listed. Concepts were considered to be common if the same concept was described in multiple papers, and during the construction of the feature model, these concepts were given a unique name. If the terminology between the papers differed, they were named according to the most commonly used term. After identifying the common and variable concepts, they were analyzed thoroughly to create distinct groups of features. This was an iterative process in which the authors held regular discussions on how the features should be structured in the model. Google Scholar was searched for identifying 15 influential papers of the IoT cybersecurity domain. Martin-Martin et al.[24] showed that, due to the wide-ranging coverage and efficient sorting algorithms, Google Scholar is a proper tool for identifying the most influential scientific papers. In order to cover the entire domain of IoT cybersecurity, a broad search query was used, namely, "Internet of Things" OR "IoT" AND "cybersecurity" OR "cybersecurity" OR "security." The papers had to be published in a peer-reviewed journal, cited over 100 times, and the content of the paper must discuss cybersecurity specific to IoT systems, services, applications, devices, technologies, networks, or a combination of these. The resulting collection of papers can be accessed from Appendix A.

Feature modeling is a popular technique for performing domain analysis.[20,24,25] A feature model of a domain represents the common and variable features of the domain.[20] By indicating the constraints and possibilities of combinations of features, the feature model describes the different models present in the domain.[25] By creating a feature model of the IoT cybersecurity domain, this research presents a visual representation of the domain, which can allow firms to create configurations of the IoT cybersecurity domain that corresponds to their needs and requirements. The top-level feature model is visualized in Figure 1. Optional features can be selected and represent a possible configuration of the domain. Relationships between features and subfeatures are represented using "Or" and "Alternative" connections. An "Or" connection indicates that one or more subfeatures should be selected in a configuration of the domain, whereas an alternative connection indicates that no more than one subfeature can be selected for a configuration. Through the usage of "Abstract" and "Concrete" characteristics, two types of relationships are represented in the feature diagram. In cases where the top-feature is a generalization of its subfeatures and cannot be implemented by itself, the top-feature is characterized as "Abstract." The "concrete" characterization is used when the top-feature can be implemented by itself. "Collapsed" indicates that there are n subfeatures underneath the top-feature. The IoT cybersecurity domain is composed of nine main features, as shown in Figure 1. These features are a high-level abstraction extracted from the 15 papers that were analyzed. Figures between 2 and 10 present the feature models of each main feature.

3.1 | Top-level feature model

The IoT cybersecurity domain is composed of nine main features, as shown in Figure 1. *Security Requirements* are the requirements that must be taken into account for security purposes throughout the lifecycle of an IoT system. *Threats and Attacks* indicate the possible threats and attacks IoT systems can face throughout their lifecycle. *Security Mechanisms* are the mechanisms that can be used to secure an IoT system against possible threats and attacks and ensure the security requirements are met. The *Security Architecture* indicates in which layers of an IoT system, the security mechanisms should be implemented. The *lifecycle* indicates the stages where security is considered necessary in the lifecycle of an IoT system, and these include design considerations for security and privacy. The *Security Strategy* features refer to the processes and policies that should be in place to ensure the secure operation of IoT systems. The *Legal Aspects* are features concerned with legislation and regulation. *Standardization* refers to the standardization efforts for IoT cybersecurity. *Application Area* features are the domains where security is considered to be essential for the IoT.

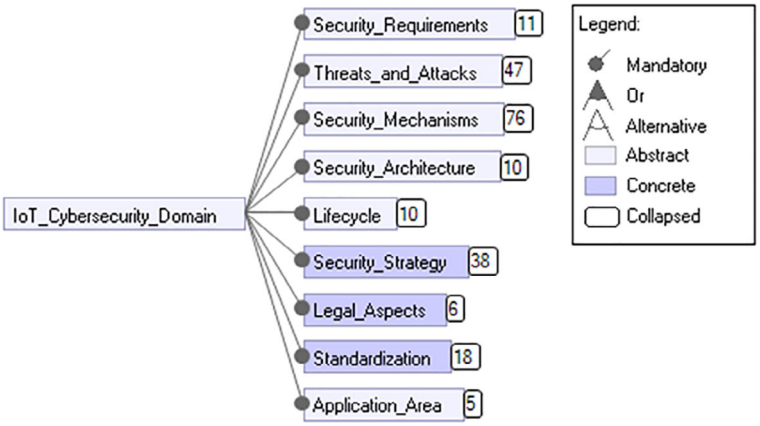

FIGURE 1 Top-level feature model of IoT cybersecurity. IoT, Internet of Things

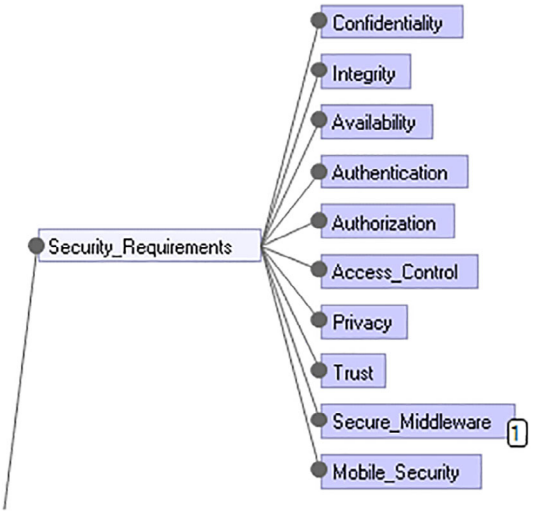

FIGURE 2 Feature model for IoT security requirements. IoT, Internet of Things

3.2 | IoT security requirements

As shown in Figure 2, these features are requirements that need to be satisfied throughout the lifecycle of IoT systems. The features *Confidentiality*, *Integrity*, and *Availability* refer to the CIA triad, these are explained in section 2.1. Availability refers to ensuring that data and devices are available to authorized users and services when they are requested. *Authentication* refers to ensuring that the devices, applications, or users requesting access to a network or requesting data are legitimate. Once a user or device in an IoT network is authenticated, it needs to be determined what its permissions are. *Authorization* is the process of identifying these permissions. *Access Control* refers to the policies or mechanisms that should be in place in order to determine who gets access to what for an IoT system. Or, in other words, it refers to the policies that, based on the authorization process, determine which resources a user or device can have access to *privacy* refers to the protection of personal information of IoT users. Information related to their location, habits, and interactions with other users should be protected and guaranteed to be safe. *Trust* is an accumulation of all the IoT cybersecurity requirements and ensures secure interactions among the different devices, layers, and applications in an IoT system. *Secure Middleware* indicates that the design and deployment of middleware should be secure. *Mobile Security* refers to the fact that all the IoT cybersecurity aspects must be met even when switching between networks or communicating over different types of networks.

3.3 | IoT security mechanisms

As shown in Figure 3, these features refer to the techniques, protocols, and other tools that can be used for securing IoT systems. All papers from the domain analysis covered some of these features, making it the most well-documented top-level feature of the domain. *Access Control Mechanisms* refer to the methods, tools, and other solutions for implementing the policies or mechanisms that control who gets access to what in IoT systems. *Authentication mechanisms* are the approaches and methods for ensuring the identification of devices and users in IoT systems. *Encryption* refers to the techniques used to encode data in such a way that only authorized users, applications, or devices can decode the data and access it. *Network*

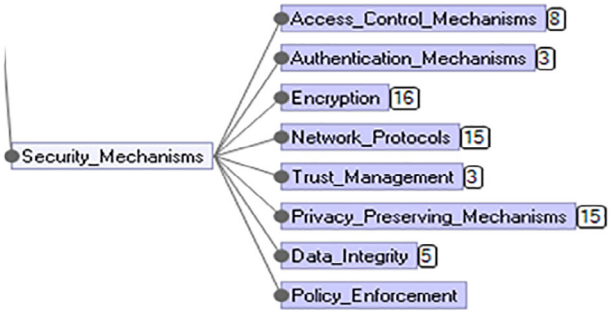

FIGURE 3 Feature model for IoT security mechanisms. IoT, Internet of Things

protocols are standards that set out the "language" that is used in an IoT system. It ensures secure communication between different devices in the system and allows devices from different manufacturers to communicate with each other. *Trust management* goes one step further than authentication and encryption, and it deals with whether the activities in the network are to be trusted rather than only relying on digital signatures and encryption keys. For example, a trust management implementation would check whether the combination of a public key and the activity its holder wants to perform can be trusted, rather than conventional mechanisms which would only check who the holder of the public key is. *Privacy-preserving mechanisms* refer to the techniques that can be implemented in an IoT system to ensure the privacy of its users and prevent data leakage to adversaries. *Data integrity* refers to methods, approaches, and techniques that can be used to ensure the integrity of data in an IoT system. *Policy enforcement* refers to the mechanisms used to enforce a given set of rules in an IoT system, and these rules are in place for the purpose of maintaining order, security, and consistency on data.[9] All these security mechanism features are mandatory as they are connected to the security requirements.

3.4 | IoT threats and attacks

In Figure 4, we present the threats and attacks an IoT system can face during its lifecycle. The features are all optional because they can happen, and the concrete features indicate protection against these can be implemented. Many of these threats and attacks are stem from traditional IT systems. The threats and attacks that are specific to the IoT are explained here. Control over devices refers to the threat of devices being controlled by adversaries to be used in criminal activities such as botnets. Physical threats are specific to IoT devices that are deployed in the environment, for example, homes, offices, factories, public spaces. These can be split up into illegal device access, where an attacker can try to extract the information they contain by physically connecting to the device, and device capture, where adversaries can capture IoT devices via physically replacing the entire device or tampering with the hardware of the node or device. Fragmentation attacks are specifically targeted against the 6LoWPAN protocol, which is a common protocol in IoT. In this type of attack, a malicious device sends forged, duplicate, or overlapping fragments of packets, that is, tiny pieces of data used for communication between computers, to other devices. Due to the little amount of available processing power, this impacts the normal functioning or the availability of IoT devices. RFID tags can be used in IoT applications, an RFID tag attack aims to trace the tag or intercept its data. Sleep deprivation attacks aim to drain the battery of IoT devices by "keeping them awake" resulting in the devices shutting down. Tag cloning is the creation of a clone of an RFID tag by a cybercriminal, because the clone is indistinguishable by the system the attacker can gain access to protected data or resources.

3.5 | IoT security architecture

The security architecture is based on Jing et al.,[26] which state that security must be ensured in all layers and across all layers. Therefore, all features are mandatory, as shown in Figure 5. Farooq et al.[27] show a similar architecture (perception layer, network layer, middleware layer, and application layer), but it is less specific than the architecture of Jing et al.[26] The perception layer refers to the security of IoT devices (i.e., nodes) and the network they have with other devices locally (i.e., the perception network), the network layer refers to the types of networks that require security. The access network is the network the IoT device is connected to, for example, Wi-Fi. The core network refers to the Internet, and the local area network is the network that connects the IoT devices to the servers where middleware is executed, that is, the application support layer. IoT application refers to the applications that run based on the data of IoT devices that are active in the field.

3.6 | IoT lifecycle

Heer et al.,[28] Miorandi et al.,[29] Roman et al.,[30] and Weber[31] covered the IoT lifecycle, as shown in Figure 6. They all state that the security of an IoT system should be guaranteed throughout its lifecycle, therefore these features are all mandatory. The naming across these four papers was not

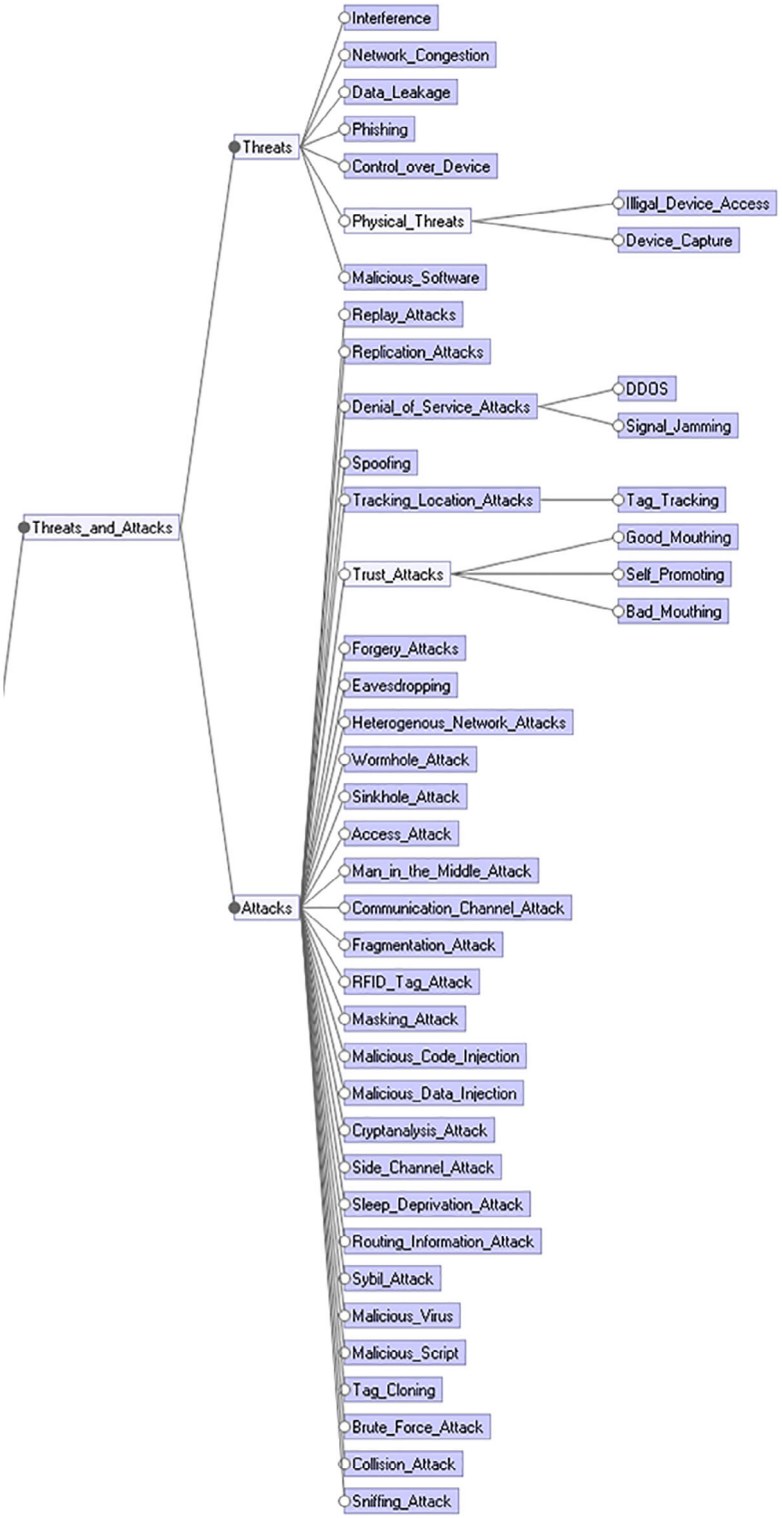

FIGURE 4 Threats and attacks an IoT system can face during its lifecycle. IoT, Internet of Things

consistent, the naming of Heer et al.[28] was deemed the most concise and thus incorporated in the feature model. Privacy and security by design are principles that ensure IoT products are secure from the moment they are available on the shelves.[5] Furthermore, during the operational phase, the IoT system should be secure, and during and after maintenance tasks (e.g., updating) they should remain secure. After the device or system has served its purpose, that is, it has reached its end-of-life, data that is privacy-sensitive must be removed[8]. In addition, during the entire lifecycle, the legal features should be taken into account, that is, the design and operation of devices should be lawful, ethical, and socially and politically acceptable.[1]

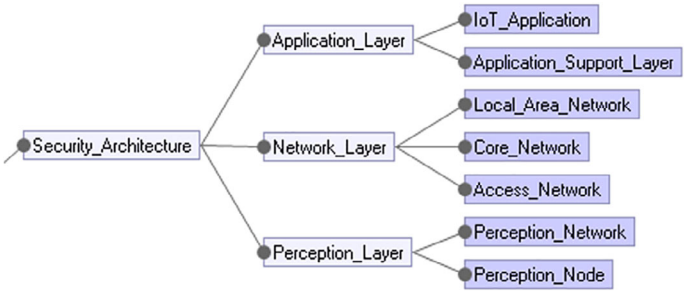

FIGURE 5 Security architecture for IoT. IoT, Internet of Things

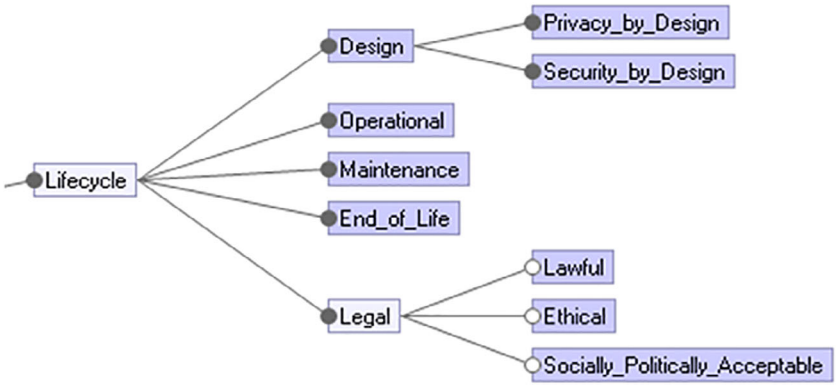

FIGURE 6 Lifecycle considerations for IoT cybersecurity. IoT, Internet of Things

FIGURE 7 Security strategy for IoT cybersecurity. IoT, Internet of Things

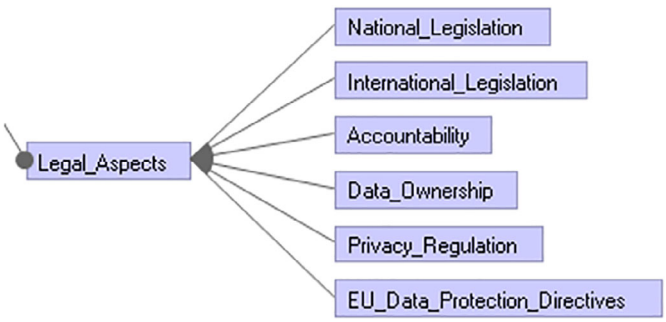

FIGURE 8 Legal aspects for IoT cybersecurity. IoT, Internet of Things

3.7 | IoT security strategies

These features refer to the processes and policies that should be in place to ensure the secure operation of IoT systems, as shown in Figure 7. These features were extracted from Bertino and Islam[32] and Farooq et al.[27]

3.8 | IoT legal aspects

These legal aspects must be considered during the lifecycle of an IoT system .[1,11] Figure 8 presents these aspects. The national, and, if applicable, international legislation must have adhered. In addition, accountability, data ownership, privacy regulations, or the EU data protection directives could be relevant. Therefore, these features were given an "Or" relationship, the relevant features depend on the region in which the IoT system is deployed.

3.9 | IoT standardization

Standardization was only covered in Weber[31] and Xu et al.[33] Figure 9 presents the subfeatures of this category. The standardization feature can be split up into three subfeatures such as goals, SO, and organization type. Goals refer to the goals of standardization, these are: to provide interoperability, compatibility, reliability, and effective operation of devices and systems.[2] Several SO are mentioned in Xu et al.'s study,[33] these are abbreviated for readability. International Telecommunication Union, International Organization for Standardization (ISO), IEEE, European Committee for Electro-technical Standardization (CENELEC), China Electronics Standardization Institute, American National Standards Institute are mentioned. However, many more standard organizations exist. Three types of organizations were identified, international SO, national SO, and legislative bodies. The relationships of "Goals" and "Standards Organizations" are "Or" relationships, as the goals of a standard do not have to be met simultaneously, and a standard can be published by one or more SO. The type of organization is either, international, national, or legislative, hence the "Alternative" relationship.

3.10 | IoT application areas

These are the domains where IoT security is considered important, based on Jing et al.,[26] Lin et al.,[34] and Xu et al.[33] They have an "Or" relationship because an IoT application is often developed for a single application area, but combinations are possible (e.g., smart thermostat for home automation and smart grid applications). Figure 10 presents these application areas.

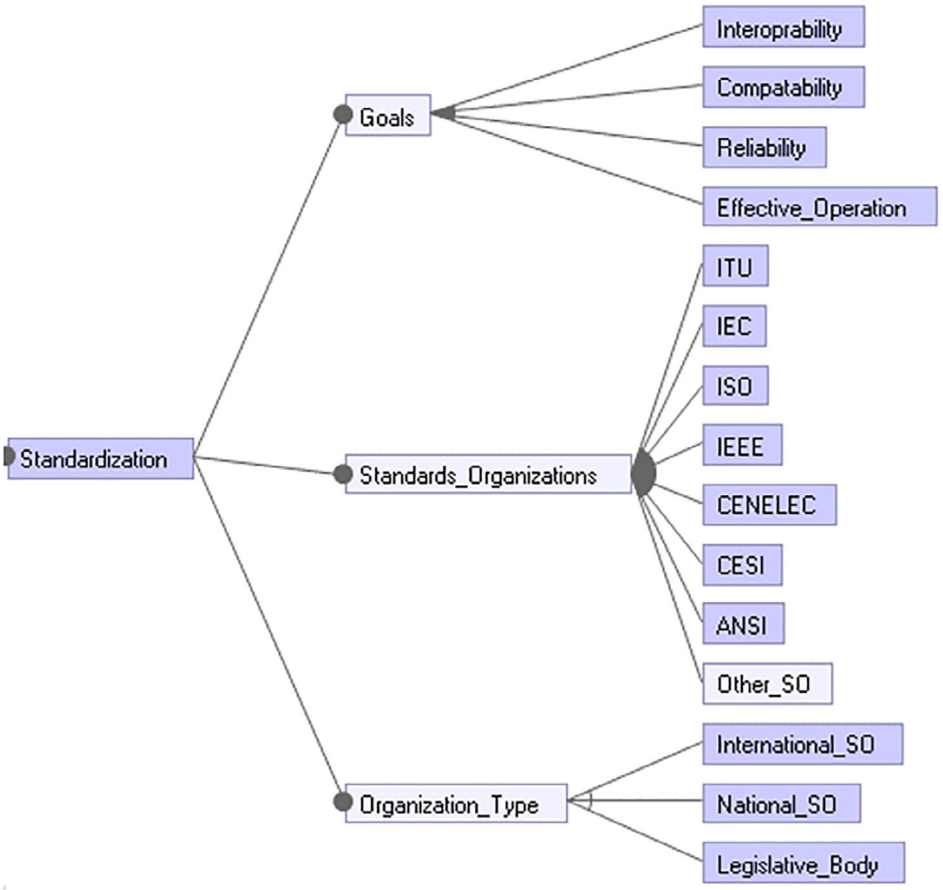

FIGURE 9 Standardization efforts for IoT cybersecurity. IoT, Internet of Things

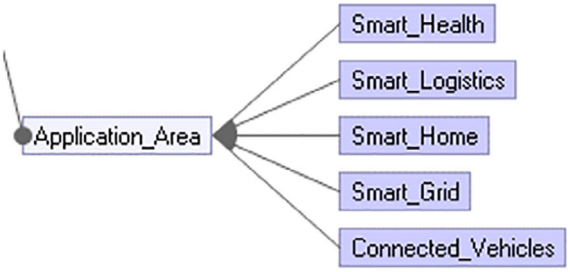

FIGURE 10 Application areas for IoT cybersecurity. IoT, Internet of Things

4 | CLASSIFICATION OF IOT CYBERSECURITY STANDARDS

4.1 | Overview of the IoT cybersecurity standards landscape

An overview of 250 unique standards was created from three bodies of research that aimed to identify relevant IoT cybersecurity standards. The studies from References 14, 18, and 22 were used to develop an up-to-date overview of the current IoT cybersecurity landscape. The overview of the collected standards first had to be cleaned. The naming of standards was split into two parts, the title of the standard and the number of the standard, as the existing overviews used them interchangeably. Sometimes they only referred to the name of the standard or only to its number. Due to the combination of multiple existing studies, the overview contained double entries, and therefore, they were removed. The status of each standard was checked, and obsolete standards were removed. For example, IETF RFC 5246, The Transport Layer Security (TLS) Protocol Version 1.2, was included in Reference 12 but was obsoleted in 2018 by IETF RFC 8446, The TLS Protocol Version 1.3. In addition, some SOs review their standards periodically; the most recent year, a standard was updated or reviewed was documented in the overview. Since IoT systems make use of multiple existing technologies, not all standards in the overview are developed specifically for the IoT. For each standard, it was checked if it was developed specifically for the IoT or not. In Figure 11, the distribution of the most recent updates is presented. From the 250 standards, 20 of them were updated in 2019, and this indicates how quickly the IoT cybersecurity standards change over time. We aimed to present this distribution explicitly because we noticed that these standards are still being developed by standard organizations. Four of the standards are under development, and this indicates that the SO is still developing.

As shown in Table 1, there are 55 unique SOs in total in the overview of which ISO/IEC has published the most standards (60). In addition, 56 of the standards from the overview were developed specifically for the IoT. The fact that 194 standards are included that are not specifically developed for the IoT indicates the number of different types of technologies that can be used in IoT systems. From the overview, the freely available standards were downloaded as PDF files. This resulted in a total of 121 standards that were available to process using NLP.

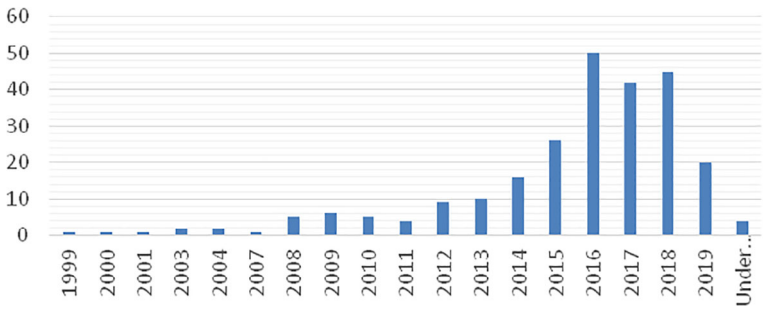

FIGURE 11 Distribution of the most recent updates of standards

TABLE 1 Distribution of the number of standards per standards organizations (SO)

Number of SOs	Number of unique standards in overview
29	1
9	2
4	3
4	4
1	6
2	8
1	11
1	15
1	16
1	25
1	26
1	60

4.2 | Classification of standards

To classify the 121 standards the following steps were completed for each standard: (1) extract the text using Textract, (2) preprocess and stem text, (3) search for each feature-query using re.findall, (4) count the amount of times the feature-query appears in the standard, (5) document the frequency of the feature-query in a Pandas DataFrame. This process results in a document vector for each of the 121 standards. Elements of the vectors represent the frequencies of the feature-queries. For example, if a document consists of the text "this is a document," and the feature-queries are "this" and "document," the result would be a vector of two elements: [1,1]. In the case of the standards, it results in a document vector of 221 elements, as the feature model consists of 221 features. The document vectors are stacked together to form a document-term matrix. This is a matrix of 121 rows (i.e., the standards) and 221 columns (i.e., the feature-queries). The standards were classified according to their three highest-scoring length-normalized TF-IDF scores, as it was considered likely that the standard covered how to implement these features. In Figure 12, we represent our classification approach. The pseudocode of the proposed method is presented in Appendix B.

4.2.1 | Extracting the text

To retrieve the features from the standard, it was decided to extract the raw text from the PDF files of the standards. The modules PyPDF2* and Textract† were applied. PyPDF2 had some difficulties with the encodings of some PDF files and returned nonalphanumerical characters rather than the true characters of the document. Textract automatically detects the encoding of a PDF file using the module Chardet‡. Hence, it was decided to use Textract with output encoding UTF-8 as this is the preferred encoding in Python 3.

After the text was extracted from the PDF file, the features could be retrieved. The feature model was converted to a CSV file, which was imported in Python as a list of strings. Each element in the list is a feature. The raw text from the standard was extracted as one long string and using Regular Expressions in Python the string was searched for each of the feature-queries.

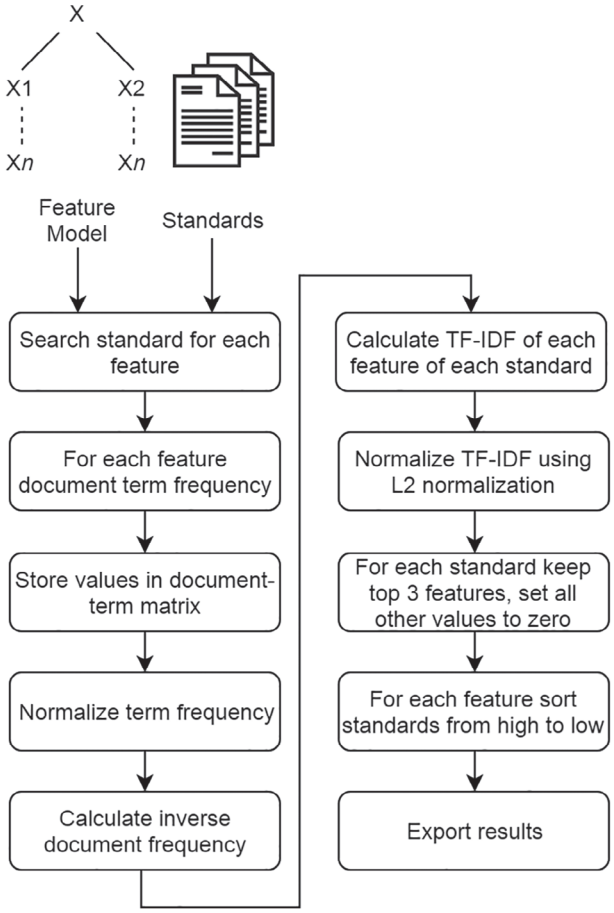

FIGURE 12 The steps from raw text to the classification

4.2.2 | Pre-processing and Stemming of Extracted Text

Numerical characters were removed to clear the text of page numbers and other numbers from tables interfering with the text.

The standards classification algorithm goes through the following preprocessing steps after the text is extracted from a PDF file:

1. Lowercase all words in the text
2. Remove newline indicators from the text, that is, \n
3. Remove all nonalphabetic characters, for example, % and 99
4. Remove all single characters from the text
5. Remove leading and trailing whitespaces
6. Remove English stop words

After the preprocessing phase, stemming was applied using the Snowball stemmer from NLTK. This results in different forms of a word being counted by the algorithm as if they are the same word.[21]

4.2.3 | Document-Term Matrix

Once the text was extracted, the feature-queries were run and the term frequencies extracted. Resulting in a document-term matrix where each row represents a standard, and each column represents a feature. The elements of the matrix were the term frequencies of the features.

The abstract features of the feature model were also included in the feature-queries because these originated from common terms in the literature. Even though these features cannot be implemented directly themselves (e.g., the lifecycle feature), it was reasoned that when standards could cover technologies or processes that are relevant to this feature, they would also mention the abstract feature. In other words, if abstract features are present in the text of a standard it indicates that the standard covers one or more subfeatures of these abstract features.

4.2.4 | TF-IDF weighting

TF-IDF can be used for numerous tasks, such as document comparison and keyword extraction. In this study, it was used for ranking and classification purposes. Term frequency is defined as the number of times term t occurs in document D (written as tf). Because the significance of a term in a text does not increase linearly with the number of times it appears in the text, the weighted version, shown in Equation (1), was used. This equation was only applied if the term frequency was not equal to 0. Thus, the features with term frequency 0 kept their original score.

$$tfw = \log_{10}(t) + 1 \quad \text{Manning et al. [21]} \tag{1}$$

Inverse document frequency (IDF) is defined as $\log_{10}$ of the total number of documents (written as n) in the corpus divided by the number of documents in the corpus that contain term t (i.e., document frequency written as $df(t)$). This was calculated using Equation (2).

$$idf(t) = \log_{10}\left(\frac{n}{df(t)}\right) \text{ Manning et al. [21]} \tag{2}$$

Because the feature selection is based upon terms from the domain analysis and from terms appearing in the documents, Lidstone smoothing was implemented based on Equation (3).

Figure 6 the steps of the classification algorithm.

$$\text{smooth } idf(t) = \log_{10}\left(\frac{1+n}{1+df(t)}\right) + 1 \text{ Manning et al. [21]} \tag{3}$$

Through the smoothing, the value one was added to the denominator, enumerator. This ensured there were no divisions by 0, which could occur because the feature-queries stem from the feature-model. That is, some features from the feature model might not appear in any of the standards from the dataset. In addition, if a feature-query occurred in all standards, the IDF would be equal to 0 (log(121/121) = 0), which would result in a TF-IDF score of 0. This implies that a standard would not be able to be classified as the feature-query. Therefore, one is also added to the result of the logarithm. By implementing smoothing, these issues were prevented. After the weighted term frequencies and the smooth IDF scores were calculated, the TF-IDF scores were calculated using Equation (4).

$$tf - idf = tf_w(t) * \text{smooth } idf(t) \text{ Manning et al. [21]} \tag{4}$$

However, the length of the standard documents differed significantly, some documents were approximately 30 pages, while others were over 3000 pages. A longer document increases the probability that the same terms occur more often, which would automatically yield a higher TF-IDF score. To be able to sort the standards from highest-ranking to lowest ranking scores, the TF-IDF scores were L_2 normalized using Equation (5). For each document, the TF-IDF score of each feature-query was divided by the square root of the sum of squares of all the TF-IDF scores of the document. By calculating the length-normalized TF-IDF scores, the feature-queries could be fairly compared with one another and be used for classification purposes.

$$tf - idf_{L2} = \frac{tf - idf}{\sqrt{(tf - idf_1^2 + tf - idf_2^2 + \ldots\ tf - idf_n^2)}} \text{ Manning et al. [21]} \quad (5)$$

4.2.5 | Steps of classification algorithm

By combining the document-term matrix that results from extracting the feature-queries from the standards, and the equations for calculating the normalized TF-IDF scores, an algorithm was designed that automatically classifies the standards according to their TF-IDF score. Running the algorithm on the 121 standards resulted in a document-term matrix of 121 by 218. Features from the feature model that were abbreviations were run as two separate queries: an abbreviated version and a written-out version. For example, CoAP was run as both CoAP and Constrained Application Protocol. The features "standardization" and "application area" and their subfeatures were not included as feature-queries in the algorithm, because it was considered unlikely that IoT cybersecurity standards would cover how to implement these features. A sample of the resulting document-term matrix with L_2 normalized TF-IDF scores is shown in Table 2.

4.2.6 | Evaluation of the algorithm

For each feature IoT system developers choose in their product configuration, the standards that cover how to implement this feature should come up. There was no validated training data available for the classification of IoT cybersecurity standards. Therefore, seven standards from the overview were randomly selected for the evaluation of the algorithm. These standards were manually classified according to features of the feature model. A standard was classified as covering a certain feature if the name of a chapter mentioned the feature or a synonym of the feature. In Table 3, an overview is presented of the results of the algorithm compared with the results of the manual classification. The columns in Table 3 indicate the following:

- *Actual*: The number of standards in the evaluation set that are manually classified as providing information about the feature and explaining how to implement the feature in an IoT system or an organization.
- *True Positive*: Feature is mentioned in the standard, and the standard explains how to implement that feature in an IoT system or in an organization.
- *False Positive*: Feature is mentioned in the standard. However, the context could be different.
- *False Negative*: Standard should be classified as a feature, but the algorithm does not classify it as expected.
- *Precision and Recall*: Precision is the fraction of correct classifications from the total number of classifications, recall is the fraction of correctly classified documents from the total number of documents in that class.

In addition to the evaluation parameters shown in Table 3, the following parameters were also calculated (Average precision rate = 0.75, Average recall = 0.72, F_1-measure = 0.73). The *F*-measure was calculated using the macro-average, giving equal weights to all the classes. This implicates that each class is as important as the other to predict correctly.[37] In addition, it should be noted that the algorithm is very efficient. It took 10 min to classify all 121 standards, of which 9.5 min were used for extracting text. For comparison, to classify the seven standards in the evaluation set by hand required a workday. A systematic approach was developed based on the experience and knowledge gained throughout this research. It reflects

TABLE 2 Sample of the document-term matrix with term frequency-inverse document frequency scores

Standard	Security requirements	Confidentiality	…	N
3GPP—TS 33.501V15.4.0	0.139	0.156	…	N
ANSI HL7—PRIVECLASSSYS, R1-2014	0	0.241	…	N
ANSI HL7—V3 IG DS4P, R1-2014	0	0.399	…	N
…	…	…	…	…
Zigbee Alliance—094945R00ZB	0.214	0.089	…	N

TABLE 3 Evaluation of the classification algorithm

Classification of feature	Actual	True Positive	False Positive	False Negative	Precision	Recall
Access control	4	2	0	2	1	0.5
Asymmetric encryption	1	0	0	1	0	0
Authentication	2	1	0	1	1	0.5
Authorization	5	4	0	1	1	0.8
CoAP	1	1	1	0	0.5	1
DTLS	0	0	1	0	0	0
ECC	1	1	0	0	1	1
Lawful	1	1	0	0	1	1
MQTT	0	0	1	0	0	0
Phishing	1	1	0	0	1	1
Policy enforcement	1	1	0	0	1	1
Privacy	2	2	1	0	0.67	1
Response to attacks	1	1	0	0	1	1
Risk assessment	1	1	0	0	1	1
RSA	1	1	1	1	0.5	0.5
Symmetric encryption	1	1	0	0	1	1
Trust	1	1	0	0	1	1

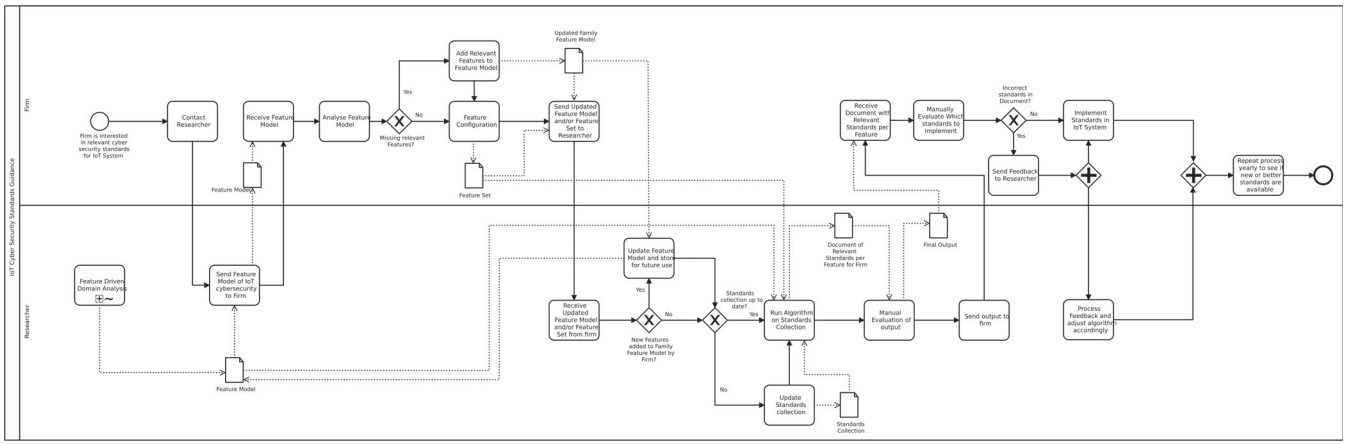

FIGURE 13 Proposed systematic approach

how the authors intend the results of this article to be used by developers of IoT systems. The systematic approach is shown in Figure 13. Various gateways were implemented where the relevancy and recency of the feature model and standards collection are checked.

5 | THREATS TO VALIDITY

The FODA was conducted using peer-reviewed papers. Including other sources, such as domain experts, could yield previously unidentified features. In addition, new features that surfaced from the standards were not added. However, the expected impact of these threats is minimal. Feature models will always evolve and can rarely be considered end-products.[23] Thus, while several features are covered through the domain analysis, there might be some features that have not been covered. However, most of the common feature categories have been included in our study. Characterization of standards could also include synonyms of the features; however, using regular synonym libraries proved difficult given the specific nature of the

standards. More delicate text extraction might improve the performance, and headers and footers might be ignored, and extra weight can be given to the chapter titles and headings. In addition, different methods, such as domain kernels might be used to build an improved classifier without needing additional training data.[38] In addition, the distribution of SO that supplied the standards in the dataset might have had an impact on the results. For example, the IETF provided 25 and ETSI 16 of the 121 standards. The terminology used specifically in these standards could impact the results. However, the IETF and ETSI[39] are large SO publishing many standards that are relevant for IoT. Hence, it is expected that this distribution of standards is an accurate representation of all available IoT cybersecurity standards. In addition, currently, there are two actors required in the business process. To ensure that the approach is as efficient, effective, and consistent as possible, an interactive tool should be created that removes the currently required sending back and forth of feature models and feature sets.

6 | CONCLUSION

This article presented a systematic approach that helps firms in selecting appropriate IoT standards for their system. To achieve this, a FODA was first conducted. This gave insights into the common and variable features of IoT cybersecurity and is the foundation of the systematic approach. The feature model allows developers of IoT systems to select the features they have or want in their IoT system. IoT cybersecurity standards presented in three previously conducted studies were combined in an aggregated overview, and metadata on the standards in the overview was collected. It was found that, in line with previous literature, the standards landscape for IoT cybersecurity is evolving quickly. In addition, it was shown that the amount of IoT-specific standards released in recent years had increased tremendously. In order to identify standards that provide information on how the features from the feature model should be implemented in an IoT system, an algorithm was developed in Python for automatically classifying IoT cybersecurity standards. Using preprocessing and term weighting techniques from the NLP domain, the algorithm calculated normalized TF-IDF scores for each feature, which was subsequently used for classifying the documents. The manual evaluation indicated that the developed method of combining feature modeling and NLP techniques for classifying standards has potential. To secure that future researchers use the feature model and algorithm in a systematic way, a BPMN diagram was presented of a proposed business process for IoT system developers who are interested in finding relevant standards for securing their IoT system. Suggestions for future research were given, including optimization of the algorithm using novel classification techniques and the development of an interactive tool that can be used by developers of IoT systems.

ACKNOWLEDGMENT

Open Access funding provided by the Qatar National Library.

CONFLICT OF INTEREST

No conflict of interest has been declared by the authors.

ENDNOTES

*https://pypi.org/project/PyPDF2/
†https://pypi.org/project/textract/
‡https://pypi.org/project/chardet/

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Bedir Tekinerdogan https://orcid.org/0000-0002-8538-7261
Cagatay Catal https://orcid.org/0000-0003-0959-2930

REFERENCES

1. ITU-T Overview of the Internet of Things; 2012.
2. Porter ME, Heppelmann JE. How smart, connected products are transforming competition. *Harv Bus Rev*. 2014;92(11):64-88.
3. Lu Y, Papagiannidis S, Alamanos E. Internet of Things: a systematic review of the business literature from the user and organisational perspectives. *Technol Forecast Soc Change*. 2018;136:285–297.
4. Porter ME, Heppelmann JE. How smart, connected products are transforming companies. *Harv Bus Rev*. 2015;93(10):96-114.
5. Alrawi O, Lever C, Antonakakis M, Monrose F. SoK: security evaluation of home-based IoT deployments. Paper presented at: Proceedings of the IEEE Symposium on Security and Privacy (SP), San Francisco, CA; 2019:208-226.
6. Computest Research: thousands of houses and offices vulnerable to hackers due to insecure domotica - Computest 21–02, 2019. https://www.computest.nl/en/news/news-and-press-releases/homes-offices-vulnerable-unsecure-domotica/. Accessed June 18, 2019.
7. Frustaci M, Pace P, Aloi G, Fortino G. Evaluating critical security issues of the IoT world: present and future challenges. *IEEE Internet Things J*. 2018;5(4):2483-2495.

8. Lyu M, Sherratt D, Sivanathan A, Gharakheili HH, Radford A, Sivaraman V. Quantifying the reflective DDoS attack capability of household IoT devices. Paper presented at: Proceedings of the 10th ACM Conference on Security and Privacy in Wireless and Mobile Networks – WiSec '17; vol. 1, 2017:46-51; ACM, New York, NY.
9. Sicari S, Rizzardi A, Grieco LA, Coen-Porisini A. Security, privacy and trust in Internet of Things: the road ahead. *Comput Netw*. 2015;76:146-164.
10. Lyytinen K, King JL. Standard making: a critical research frontier for information systems research. *MIS Q*. 2006;30:405-411.
11. Li S, Da Xu L, Zhao S. The Internet of Things: a survey. *Inf Syst Front*. 2015;17(2):243-259.
12. Hogan M, Piccarreta B. *Interagency Report on the Status of International Cybersecurity Standardization for the Internet of Things (IoT)*. Gaithersburg, MD; National Institute of Standards and Technology; 2018.
13. Silverthorne V. What to expect from the fluid IoT standards landscape in 2019 | Page 2; Februray 02, 2019. [Online] https://www.iotworldtoday.com/2019/02/14/iot-standards-in-2019-semantics-security-and-social-issues/2/. Accessed. June 18, 2019.
14. Brass I, Tanczer L, Carr M, Elsden M, Blackstock J. Standardising a moving target: the development and evolution of IoT security standards. *Living in the Internet of Things: Cybersecurity of the IoT - 2018*; UK: Institution of Engineering and Technology; 2018:24 9 pp.
15. Schatz D, Bashroush R, Wall J. Towards a more representative definition of cyber security. *J Dig Forens Secur Law*. 2017;12(2):53-74.
16. Baron J, Spulber DF. Technology standards and standard setting organizations: introduction to the searle center database. *J Econ Manag Strateg*. 2018;27(3):462–503.
17. Hogan M, Newton E. *Supplemental Information for the Interagency Report on Strategic*. Gaithersburg, MD: U.S. Government Engagement in International Standardization to Achieve U.S. Objectives for Cybersecurity; 2015.
18. Riahi Sfar A, Natalizio E, Challal Y, Chtourou Z. A roadmap for security challenges in the Internet of Things. *Dig Commun Netw*. 2018;4(2):118-137.
19. Keoh SL, Kumar SS, Tschofenig H. Securing the Internet of Things: a standardization perspective. *IEEE Internet Things J*. 2014;1(3):265-275.
20. Lee K, Kang KC, Lee J. Concepts and guidelines of feature modeling for product line software engineering. In: Gacek C, ed. . Vol 2319. Berlin/Heidelberg, Germany: Springer; 2002:62-77.
21. Manning CD, Raghavan P, Schütze H. *An Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press; 2009.
22. Manning CD, Raghavan P, Schütze H. Vector space classification. *Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press; 2009:289-317.
23. Kang KC, Cohen SG, Hess JA, Novak WE, Peterson AS. *Feature-Oriented Domain Analysis (FODA) Feasibility Study. Technical Report CMU/SEI-90-TR-222*. Pittsburgh, PA: Software Engineering Institute, Carnegie-Mellon University; 1990.
24. Martin-Martin A, Orduna-Malea E, Harzing A-W, Delgado López-Cózar E. Can we use Google scholar to identify highly-cited documents? *J Inform*. 2017;11:152-163.
25. Tekinerdogan B, Öztürk K. Feature-driven design of SaaS architectures. *Software Engineering Frameworks for the Cloud Computing Paradigm*. London, UK: Springer; 2013:189-212.
26. Jing Q, Vasilakos AV, Wan J, Lu J, Qiu D. Security of the Internet of Things: perspectives and challenges. *Wirel Netw*. 2014;20(8):2481-2501.
27. Farooq MU, Waseem M, Khairi A, Mazhar S. A critical analysis on the security concerns of Internet of Things (IoT). *Int J Comput Appl*. 2015;111(7):1-6.
28. Heer T, Garcia-Morchon O, Hummen R, Keoh SL, Kumar SS, Wehrle K. Security challenges in the IP-based Internet of Things. *Wirel Pers Commun*. 2011;61(3):527-542.
29. Miorandi D, Sicari S, De Pellegrini F, Chlamtac I. Internet of Things: vision, applications and research challenges. *Ad Hoc Netw*. 2012;10(7):1497-1516.
30. Roman R, Zhou J, Lopez J. On the features and challenges of security and privacy in distributed Internet of Things. *Comput Netw*. 2013;57:2266-2279.
31. Weber RH. Internet of Things – new security and privacy challenges. *Comput Law Secur Rev*. 2010;26(1):23-30.
32. Bertino E, Islam N. Botnets and Internet of Things security. *Computer (Long Beach Calif)*. 2017;50:76–79.
33. Da Xu L, He W, Li S. Internet of Things in industries: a survey. *IEEE Trans Ind Inform*. 2014;10(4):2233-2243.
34. Lin J, Yu W, Zhang N, Yang X, Zhang H, Zhao W. A survey on Internet of Things: architecture, enabling technologies, security and privacy, and applications. *IEEE Internet Things J*. 2017;4(5):1125-1142.
35. Manning CD, Raghavan P, Schütze H. Text classification and Naive Bayes. *Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press; 2009:253-287.
36. Haddi E, Liu X, Shi Y. The role of text pre-processing in sentiment analysis. *Proc Comput Sci*. 2013;17:26-32.
37. Kay M, Patel SN, Kientz JA. How good is 85%? a survey tool to connect classifier evaluation to acceptability of accuracy, Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015, pp. 347–356.
38. Gliozzo A, Strapparava C. Domain kernels for text categorization. Paper presented at: Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL), Ann Arbor, Michigan; 2005:56-63.
39. ETSI SmartM2M: IoT standards landscape and future evolutions (TR 103 375 - V1.1.1); 2016.
40. Zhou L, Chao H-C. Multimedia traffic security architecture for the Internet of Things. *IEEE Netw*. 2011;25(3):35-40.
41. Granjal J, Monteiro E, Sa Silva J. Security for the Internet of Things: a survey of existing protocols and open research issues. *IEEE Commun Surv Tutor*. 2015;17(3):1294-1312.
42. Bandyopadhyay D, Sen J. Internet of Things: applications and challenges in technology and standardization. *Wirel Pers Commun*. 2011;58(1):49-69.
43. Kothmayr T, Schmitt C, Hu W, Brünig M, Carle G. DTLS based security and two-way authentication for the Internet of Things. *Ad Hoc Netw*. 2013;11(8):2710-2723.

How to cite this article: van der Schaaf K, Tekinerdogan B, Catal C. A feature-based approach for guiding the selection of Internet of Things cybersecurity standards using text mining. *Concurrency Computat Pract Exper*. 2021;33:e6385. https://doi.org/10.1002/cpe.6385

APPENDIX A. SELECTED PAPERS FOR DOMAIN ANALYSIS

Author	Year	Title	Cited	Journal
Weber[31]	2010	Internet of Things – New security and privacy challenges	968	Computer Law and Security Review
Xu et al.[33]	2014	Internet of Things in industries: a survey	1758	IEEE Transactions on Industrial Informatics
Sicari et al.[9]	2015	Security, privacy and trust in Internet of Things: the road ahead	738	Computer Networks
Jing et al.[26]	2014	Security of the Internet of Things: perspectives and challenges	553	Wireless Networks
Roman et al.[30]	2013	On the features and challenges of security and privacy in distributed Internet of Things	587	Computer Networks
Zhou and Chao[40]	2011	Multimedia traffic security architecture for the Internet of Things	278	IEEE Network
Granjal et al.[41]	2015	Security for the Internet of Things: a survey of existing protocols and open research issues	385	IEEE Communication Surveys
Heer et al.[28]	2011	Security challenges in the IP-based Internet of Things	258	Wireless Communications
Lin et al.[34]	2017	A survey on Internet of Things: architecture, enabling technologies, security and privacy, and applications	318	IEEE Internet of Things Journal
Li et al.[11]	2015	The Internet of Things: a survey	713	Information System Frontiers
Bandyopadhyay and Sen[42]	2011	Internet of Things: applications and challenges in technology and standardization	876	Wireless Personal Communications
Bertino et al.[32]	2017	Botnets and Internet of Things security	126	Computer
Miorandi et al.[29]	2012	Internet of Things: vision, applications and research challenges	2350	Ad Hoc Networks
Farooq et al.[27]	2015	A critical analysis on the security concerns of Internet of Things (IoT)	147	International Journal of Computer Applications
Kothmayr et al.[43]	2013	DTLS-based security and two-way authentication for the Internet of Things	211	Ad Hoc Networks

APPENDIX B. PSEUDOCODE OF THE PROPOSED METHOD

```
USERINPUT
    feature_family = IOT_CyberSecurity_Features.csv
    feature_set = Feature_set.csv
    output_dir = Analysed
    stemmer = english
    complete_stopwords = english
    top_k = 5

put_feature_model_and_feature_set_into_a_list
#Transform each feature to lowercase, remove leading and trailing
#spaces from every string in the list,
#and replace with whitespace
clean_up_feature_list
remove_stopwords_and_stem_features
fix_issues_created_by_stemming
count_the_number_of_times_a_feature_mentioned
extract_clean_text_from_PDF_files_in_folder
normalize_raw_wordcount_and_calculate_normalized_TF-IDF
classify_according_to_tf-idf_weighting_and_identify_top_k_standards
export_the_tf-idf_matrix_and_the_classification_of_standards
```