algorithms

MDPI

Article

Comparative Analysis of Classification Methods and Suitable Datasets for Protocol Recognition in Operational Technologies

Eva Holasova *, Radek Fujdiak * and Jiri Misurec

Department of Telecommunications, Faculty of Electrical Engineering and Communication, Brno University of Technology, Technicka 12, 616 00 Brno, Czech Republic; misurec@vut.cz
* Correspondence: eva.holasova@vut.cz (E.H.); fujdiak@vut.cz (R.F.)

Abstract: The interconnection of Operational Technology (OT) and Information Technology (IT) has created new opportunities for remote management, data storage in the cloud, real-time data transfer over long distances, or integration between different OT and IT networks. OT networks require increased attention due to the convergence of IT and OT, mainly due to the increased risk of cyber-attacks targeting these networks. This paper focuses on the analysis of different methods and data processing for protocol recognition and traffic classification in the context of OT specifics. Therefore, this paper summarizes the methods used to classify network traffic, analyzes the methods used to recognize and identify the protocol used in the industrial network, and describes machine learning methods to recognize industrial protocols. The output of this work is a comparative analysis of approaches specifically for protocol recognition and traffic classification in OT networks. In addition, publicly available datasets are compared in relation to their applicability for industrial protocol recognition. Research challenges are also identified, highlighting the lack of relevant datasets and defining directions for further research in the area of protocol recognition and classification in OT environments.

Keywords: classification methods; datasets; machine learning; operational technology; protocol classification; protocol recognition; security

check for updates

Citation: Holasova, E.; Fujdiak, R.; Misurec, J. Comparative Analysis of Classification Methods and Suitable Datasets for Protocol Recognition in Operational Technologies. *Algorithms* **2024**, *17*, 208. https://doi.org/10.3390/a17050208

Academic Editors: Nuno Fachada and Nuno David

Received: 31 March 2024
Revised: 8 May 2024
Accepted: 9 May 2024
Published: 11 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cyber security is now an essential part of industrial networks. As a result of the interconnection of Operational Technology (OT) and Information Technology (IT), new possibilities for remote management, the use of cloud storage, real-time data transfer over long distances, or integration between different OT and IT networks, for example, are emerging. On the other hand, there are new security risks to which OT networks are exposed [1]. OT networks used to be completely isolated from IT networks, so there was not much emphasis on cyber security [2]. For this reason, there is a new emphasis on monitoring and analyzing OT traffic.

Protocol recognition and classification is an important task in security control and can be conducted via data analysis [3]. Knowledge of the protocols used in the network contributes to network optimization and helps to understand how traffic is distributed and what data are present in the network. Based on protocol recognition and data classification, traffic routes can be optimized, the quality of traffic and transmitted data can be improved, and network management strategies can be developed. Based on the automatic inspection of traffic data, redundant messages can be filtered, and the volume of transmitted messages can be reduced, thereby reducing the computational complexity and cost of transmission. In terms of network security, the use of protocol recognition leads to earlier and timely detection of threats, for example, in the case of a Man in the Middle attack. It is also possible to detect and find a virus early. There is a large number of methods that can be used to achieve protocol identification both in IT and OT networks. It is possible

to use traditional methods, which include classification based on ports used, or more sophisticated approaches using Artificial Intelligence (AI). Using such approaches, it is possible to perform an in-depth analysis of the monitored data stream (or other data units) and classify not only the protocol used but also, for example, the cipher suite used. Performing classification in OT networks is currently less common, but it provides great potential in terms of security benefits for such networks. It is for this reason that this paper has been created, in order to describe and summarize the different protocol classification methods (especially in OT networks) and also the available datasets.

This paper focuses on the protocol recognition aspects of OT networks. Advanced methods using AI techniques can be used to perform protocol recognition with additional recognition capability. Conventional techniques, such as relying on known ports, may not be fully sufficient and thus more advanced techniques that are able to directly detect/recognize the protocol itself (trained marks of the protocol) need to be employed. Based on the analysis of the current state of the art, it is clear that the recognition of industrial protocols is rather minor, as is the current state of publicly available datasets. Thus, this paper points out this gap (research gap), and for this reason, it performs (i) a summarization of methods for network traffic classification, (ii) a summarization of methods for recognition and identification of the protocol used in the network, (iii) the use of machine learning methods for industrial protocol recognition. Finally, (iv) an analysis of publicly available datasets that can be used for industrial protocol classification was performed. This article takes aim at the scientific question: How can industrial protocol classification be achieved? What publicly available datasets can currently be used specifically for the purpose of classifying these protocols? OT networks require increased attention due to IT and OT convergence, in particular, due to the increased risk of cyber-attacks that may target these networks. Convergence has caused, among other things, a proliferation of attack vectors, making advanced data monitoring necessary and using a software-as-a-service (SaaS) approach. Industrial protocol classification thus enables (i) automatic detection of the protocol used to assess security, including the cipher suite used, (ii) diagnostic data, network monitoring (protocol usage within different sectors, etc.), (iii) automation of audit tools, and (iv) development of protocol adaptive solutions—automatic protocol detection and further actions following this knowledge.

The structure of this paper is as follows: Section 2 describes the specifics of OT networks, the effects of the convergence of IT and OT networks, and, hence, the need to use sophisticated methods to enhance security in the OT industry. Section 3 presents an analysis and comparison of the current state of the art, focusing mainly on the issues of protocol recognition and traffic classification. Furthermore, Section 4 presents various methods for the purpose of traffic analysis. The section presents approaches to protocol classification, recognition and identification, Machine Learning (ML) methods, and metrics used to evaluate models. Section 5 focuses on the available datasets usable for the purpose of traffic classification and the chosen protocol, and a comparison of the most relevant datasets in terms of several parameters is also provided.

2. Operational Technology Networks Specifics

A significant difference between classical IT networks and OT networks is their purpose and related use. OT networks have the main purpose of controlling and monitoring the industrial process, whereas IT networks aim mainly at data transmission (by nature non-critical in comparison with OT networks). Another distinction is the elements and components of the individual networks themselves. IT networks use end stations (laptops, desktop PCs, mobiles, tablets, etc.). Network elements and infrastructure provide data transfer mainly between end-user elements using data stored on servers (located and connected to the Internet). On the other hand, OT networks typically use specific devices with a well-defined purpose to provide/monitor a specific activity within an industrial process. These can be single active/passive elements (actuators and sensors), control PLCs, HMIs (providing visualization of the current process status to the operator), or SCADA/DCS

components. Another difference is the data transmitted itself and the typical orientation of the data flow. Within IT, it is mainly the use of data obtained from higher layers (Internet) and its local modification/processing/consumption by the user. OT networks mainly generate data from sensors and perform operations by actuators. Thus, the data occurring in an industrial network mainly contains data acquired from sensors (temperature, pressure, speed—numerical data), and based on these data, actuators (motors, pumps, valves—binary state I/O) are activated/deactivated [4].

Another major difference is the security of individual networks. IT networks are evolving at a very fast pace, the lifetime of equipment within IT networks is typically 3–5 years (servers, workstations, laptops and network components), and the frequency of updates is also very high. Systems and software are regularly updated and upgraded to improve performance, security, and functionality. In terms of basic requirements, the priority is security, i.e., confidentiality of data, followed by data integrity, with availability (CIA triad) coming in third. Thus, it is necessary to transfer the data primarily in a confidential manner (encryption), ensuring their integrity (preserving the content without modification), followed by their availability (slight delays and outages are tolerated to provide more critical services—there is no security risk not to deliver the message immediately). In contrast, OT networks are completely identical in these aspects. Development within OT is slow and gradual, equipment lifetimes are typically 10–20 years (i.e., decades), so the frequency of updates is conducted at large intervals (industrial process is affected—creating downtime and slowing production efficiency). Systems often require long-term stability and reliability, which means that updates or changes are made less frequently to avoid the risk of disrupting critical operations [4].

In terms of basic requirements, the priority is data and service availability, followed by integrity and thirdly confidentiality (AIC triad). It is, therefore, necessary to have available data from the industrial process at all times to be able to monitor and manage the process adequately and in a timely manner. This is important because of the nature of OT networks, where critical parts are controlled and where there is a risk of malfunction or danger to human health in the event of a process disturbance (nuclear power plants, thermal power plants, etc.). It is also necessary that data integrity is preserved, and only after these tears are preserved is the safety considered. IT and OT networks differ in the nature of the services they provide in terms of their importance. They differ in terms of priorities and especially in terms of the security of the data transmitted. They also differ in the subject/scope of the data transmitted. They also differ in the individual elements of the network. Another difference is the upgrades performed, where OT is significantly more complex than IT, as well as the replacement/upgrade of equipment (OT requires a higher lifetime).

2.1. Information Technology and Operational Technology Convergence

IT and OT convergence represent the current trend of interconnection of individual components, especially their availability via the Internet. This convergence involves the integration of existing OT networks and structures within the IT network. This convergence facilitates the use of the current trend of software as a service, especially for the processing and evaluation of available data, where this was often not possible before, and data could not leave the closed and isolated network. It is equally possible to remotely access and manage these data. While this brings a number of benefits, it also involves challenges that need to be addressed, particularly from the security perspective. The problem is the long-term enclosure of OT infrastructures, which has ensured security in terms of physical security. In order to access the assets, it was necessary to overcome physical security, and only then could the assets be accessed. It is convergence, however, that significantly alters this approach. There is no need to overcome physical security, and it is possible to access assets from a SW perspective without breaching the security perimeter (from a physical perspective).

Convergence increases the risk of a security incident compared to a closed approach [2]. Due to the long-term closed nature and reliance on physical security alone, security mecha-

nisms and protocols are not at the same level as in IT networks. OT networks use industrial protocols for the transmission of individual data, which are specific protocols tailored for the transmission of sensor data and individual commands. However, these protocols often do not support confidentiality and integrity, and thus, various extensions and additional mechanisms have to be used. Thus, from a software perspective, OT networks represented insecurity by design. The challenges involved are to ensure the security of the unsecured protocol in such a way that the priorities of the requirements (availability first) are not affected. Thus, it is not possible to use current mechanisms from IT networks and apply them directly to the OT network environment without modification. Similarly, a secure separation of the IT and OT network must be implemented in such a way that the OT network is maximally separated from the rest of the network. This requires the use of firewalls, DMZ, IDS, and IPS mechanisms in conjunction with AI-enabled applications.

The impacts of the convergence of IT and OT networks include a significant proliferation of attack vectors. Convergence has made these networks "accessible" to the attacker, and physical security is no longer the main security measure. Thus, it is now (from a cyber-security perspective) a basic block that is as necessary as it used to be but no longer represents the main attack vector. Attackers can exploit the very interface that makes the connection between IT and OT networks. In particular, this may include internal services for managing and monitoring industrial processes [5]. In conjunction with these systems and devices, in general, within OT networks there are passphrases and inbuilt security measures. Insufficient quality/complexity of passphrases and excessive system measures also degrade cyber-security. The human factor is also a risk, especially in terms of social engineering or phishing attacks. According to [5], the first place in the attack vector is the compromise of IT systems, followed by the use of engineering workstations, and the third place is external remote services.

Protocol classification will help, especially with protocol security checks in the form of internal audits, etc. Knowledge of the protocols will also help with diagnostic data and obtaining an overview of the traffic occurring within the monitored OT networks. Finally, the development of a good industrial protocol classification method will help with the development of new devices. It is the automatic protocol recognition that will enable the creation of devices that automatically recognize the protocol in the network and can use this knowledge to, for example, automatically inspect and set firewall rules. In order to best secure the OT network, these approaches need to be combined. It is necessary to use tools for detecting security incidents, classifying the protocols used, as well as educating the human factor. The emergence of automated tools would enable effective control (auditing) and also the supervision of critical network elements. Industrial protocol classification can be used at individual industrial facilities (factories, plants, etc.), but also within various SaaS service providers, which can monitor and classify traffic within the network. Last but not least, this method can be used to perform non-invasive security checks of industrial protocols without the need to access the data themselves directly. It is thus possible to use the encrypted form of the messages and to perform protocol classification on this basis, including its cipher suite.

The early detection of security incidents helps to activate adequate countermeasures. Using the knowledge of the type of anomaly, it is possible to activate appropriate countermeasures so that the impact on the industrial process itself is minimized. This is related to the critical nature of the industrial processes themselves, where system shutdown can mean potential damage. It is thus advisable to perform a timely, safe system shutdown. However, the aim is to prevent such safety incidents. To do just that, it is advisable to use industrial protocol classification in the form of a security audit and implement appropriate countermeasures to minimize the likelihood of a security threat.

2.2. Operational Technology Hierarchy Model

The individual physical operations and related control and monitoring components are sorted according to IEC 62443 [6] (also known as the Purdue model) into individual

layers (six in total). This division is made according to the purpose of each layer so that individual operations can be scaled and safety levels defined within the manufacturing process. Communication within the model is vertical between the layers to achieve effective control and monitoring of the process. The layers closer to the product itself (processed through the L0 layer) form the core and basic building blocks of OT networks. As the layers grow, they gain abstraction and gradually move into the IT network. The individual layers contain differently sensitive information, and therefore, it is necessary to maintain an adequate trust level (preferably zero-trust) [7]. A graphical visualization of such a model is shown in Figure 1, which shows the Purdue model as well as the basic blocks from the RAMI 4.0 model (right side). Level 0 (Field level) contains components directly dedicated to control, the actual execution of an activity using sensors (getting values) and actuators (executing activities).

Data sent to/from L0 is conducted from the L1 layer (Control level). This layer contains the individual Programmable Logic Controller (PLC), Distributed Control System (DCS), and PID devices. These are the components that acquire data from sensors and actuators (simplified as the first logic unit that evaluates the acquired data and can convert them into digital form). These units directly control the process through the connected actuators. The decision to intervene in the process can be initiated directly from L1 or by devices from L2 (Supervisory level) [4].

Within L2 there are parent PLCs that collect data from the slave PLCs and make process modifications based on the defined operations/schemes and settings. This layer (L2) also houses workstations (operator/attendant workstations) and the local Human Machine Interface (HMI), which is used to display the current status to the operator. The process can thus be controlled via the HMI, workstations, or status evaluation by the supervisor PLC from L2, then the data are passed to the PLCs on the L1 level, and they trigger the required actions on L0.

The fourth layer (L3—Planning level) serves mainly as a support layer for the whole system. Global HMI and other server services can be located within this layer. This may include Dynamic Host Configuration Protocol (DHCP), Domain Name System (DNS), Lightweight Directory Access Protocol (LDAP), and Network Time Protocol (NTP) servers. In addition, historian servers are often located at this level to provide specific services such as storing historical data (describing the behavior and state of the process over time), analyzing stored values, and archiving events/process states over time. This layer also contains Supervisory Control and Data Acquisition (SCADA) or DSC.

Both systems are used for data acquisition from the OT network and process control. The main objective of SCADA is data acquisition; networks consist of multiple Remote Terminal Units (RTUs) that are used to collect data back to the central control system where they can be used to make higher-level decisions (based on a global view of the data). DCS is mainly used for on-site process control, connecting PLCs, sensors/actuators state, and workstations. The main objective is to collect data and control the process from devices located closer to L0. The main difference between DCS and SCADA is, therefore, in their focus and application. DCS is more focused on automating and controlling manufacturing processes within a single facility or complex, while SCADA focuses on monitoring and controlling equipment spread over large areas with an emphasis on data collection and surveillance.

Layer L4, as well as L5 can be referred to as the management level. L4 is used to provide scheduling and provisioning of other local services (e.g., printing, web server, or domain controller), so it is the Plant operational level. This layer can also contain a historian mirror and a remote access server. In general, Manufacturing Execution Systems (MES) are software solutions that actively improve the quality and efficiency of manufacturing processes.

The L5 layer focuses on enterprise applications and Enterprise Resource Planning (ERP). However, the L4 and L5 layers are very intertwined.

This model can also be supplemented with a layer that vertically connects all the layers. This concept is referred to as NAMUR Open Architecture (NOA) [8]. The aim is to enable secure, flexible, and efficient interconnection of OT with IT without compromising the functioning of critical process control systems. This may involve the collection of data from additional sensors located on the equipment. Where these devices cannot directly compromise the process itself (there is no direct connection between the sensors and the OT infrastructure), there is a one-way data flow from OT to IT.

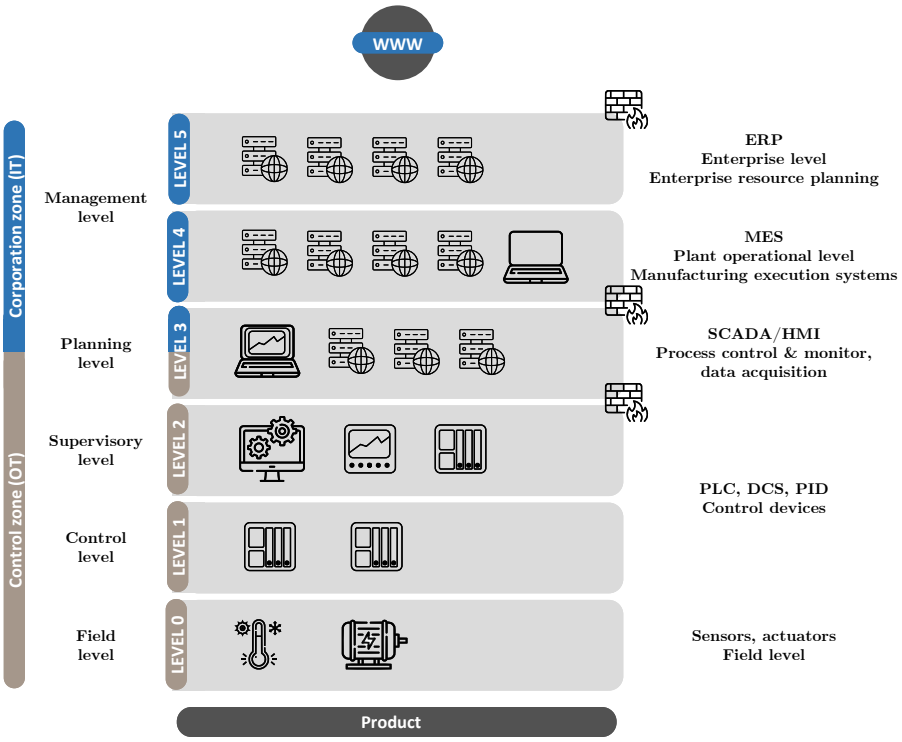

Figure 1. Hierarchical structure within OT networks expressed through the Purdue model.

For completeness, it should be noted that the term OT refers to hardware and software that directly monitors and controls physical equipment, processes and events in an industrial environment. OT includes Industrial Control Systems (ICS), which are specifically designed to control and automate industrial processes. ICS includes a variety of systems, including SCADA or DCS. OT networks have different requirements compared to IT networks. This is due to the nature of these networks and, in particular, their purpose. In the development of these networks, it is necessary to use up-to-date approaches such as ML and NN techniques, both for the detection of security incidents (traffic classification) and for the recognition and identification of the industrial protocols used. The convergence of IT and OT networks is putting pressure on the security of these networks, but it is always necessary to consider the appropriateness of individual measures in such a way that the functionality of the OT networks themselves is not compromised. The use of ML and NN techniques has the potential to enhance the security of OT networks and, in particular, can be used in such a way that they do not cause additional load to these networks. If used appropriately, a non-invasive way of using the available data can be achieved.

3. State of the Art

Protocol classification provides benefits, especially for automatic processing and automatic monitoring of data on the network. The use of classification in OT brings the benefits of enabling the development of protocol-independent approaches, especially in the area of cybersecurity. Therefore, it enables the automated management of data flows, the creation and modification of detection and mitigation rules, etc. Table 1 shows an

overview of the current approaches to protocol recognition and traffic classification in both IT and OT industries. In general, supervised approaches, e.g., machine learning and neural networks, are required in classification. A common approach is the use of convolutional neural networks, where data streams, frames, or other data structures are visualized into image data and these are then identified through convolutional neural networks.

In general, traffic classification is also more common than protocol recognition. Protocol classification can be more challenging than traffic classification (this is evident from the success rates achieved by the models). Performing protocol classification in the OT sector is particularly important in the case of encrypted traffic. In the case of encryption, it is not possible to use common (generic) protocol identification methods, such as known port recognition at the transport layer level, or to use multiple parsers to find a match. Due to IT and OT convergence, it is also necessary to assume different masking techniques performed by the attacker, also for this reason, these classification methods are very important. Similarly, in the case of protocol recognition in OT, it is possible to recognize not only the industrial protocol itself but also other parameters, such as the type of cipher suite chosen.

In total, a comparison of 20 different approaches is made, where protocol recognition in OT networks is only addressed in a minimum of current literature, and most of them target IT networks. In the case of traffic classification, the ratio is more balanced. In the case of protocol recognition, OT networks are particular and present a significant challenge due to their distinct differences. Similarly, a small number of publicly available datasets focus on this issue. Finally, it often relies only on selected ports at the transport layer level. AI methods are not used in the case of protocol recognition in OT networks, even though these methods can represent a great cyber benefit (especially in connection with Industry 4.0+). A large number of works have focused on traffic classification in IT and OT networks. Most of the works focus on cyber-security with the aim of network anomaly detection/classification. This approach (traffic classification) thus represents the implementation of a classification of the data transmitted inside a chosen traffic protocol.

For classification reasons, a supervised approach is generally used, often in combination with convolutional neural networks (CNNs). This approach represents a method in which data blocks are expressed using visual representation, i.e., the conversion of information into image data. This may be processing at the level of data streams, packets, or other data units. Some papers also focus on the encrypted data stream (encoding column). This area presents great potential from the cybersecurity perspective, where it is possible to perform traffic recognition without having to decrypt the traffic. This can be particularly beneficial when processing large amounts of data, for example, at the network administrator level or for the purpose of monitoring whether industrial data are leaving specified sections. Also, most approaches do not focus on real-time classification, but delay-independent classification is performed. It is the low delay in the classification performed that allows the use of these methods (protocol recognition, traffic classification) in the control mechanisms performing the classification of the actual network traffic. Often, authors do not provide datasets, so the classification of the protocol or network traffic is performed on a dataset that is not publicly available. Thus, it is not possible to re-evaluate the results, directly relate the results to the obtained results, or compare different approaches for classification purposes. Custom (own) datasets that are no longer available bring significant limitations in the development and comparison of available tools and approaches.

Based on the analysis of the current state of the art, the main challenges can be identified as (i) the creation of suitable and publicly available datasets that are oriented towards industrial protocols. These datasets must also contain multiple industry protocols in order to validate the discriminative capabilities of each approach. Furthermore, (ii) focusing on the potential in the area of encrypted traffic (protocols) in OT networks. (iii) Comparing the different processing approaches of the developed dataset and identifying the main research direction.

Table 1. Comparison of relevant literature in protocol recognition and traffic classification from the perspective of IT and OT infrastructures.

Methods	Type	Year	Technique	Model	ML Type	Protocols	Encoding	Real-Time	Epochs	Layers	Accuracy [%]	Datasets	Ref.
Protocol recognition	OT	2021	CNN	AM-ADCNN + LSTM	Supervised	4	No	No	20	-	93.0	Own	[9]
		2023	DNN	PREIUD	Unsupervised	1	No	No	-	-	-	Own	[10]
	IT	2011	Network Packet Inspection	Deterministic Finite-state Automaton	-	9	No	No	-	-	-	Own	[11]
		2012	Fingerprinting	-	-	4	No	Yes	-	-	95.0	Own	[12]
		2017	CNN + RNN	CNN + RNN-2a	Supervised	15	No	No	60–90	9	99.6	RedIRIS [13]	[14]
		2020	CNN	PtrCNN	Supervised	4	No	Yes	20	8	96–100	DARPA [15]	[16]
		2020	CNN	-	Both	3	No	No	-	-	75.8–89.8	Own	[17]
		2021	Pattern matching algorithm	-	Supervised	4	No	No	-	-	93.8–100	DARPA [15]	[18]
		2021	CNN	ICLSTM	Supervised	12	Yes	No	-	-	97.5	ISCX 2016 [19]	[20]
		2023	CNN	-	Supervised	8	Yes	No	-	-	98.2	ISCX VPN-nonVPN [19]	[21]
Traffic classification	OT	2019	ML	DT, KNN, SVM, NB	Supervised	1	Yes	No	-	-	95.0	Own	[22]
		2019	Traffic Fingerprinting	CART	-	-	No	No	-	-	94.8	SWaT [23], SCADA Network Data Sets for Intrusion Detection Research [24]	[25]
		2020	ML	KNN, SVM, DT, NBG, BKNN, BT, RF, AdaBoost, GB	Both	1	No	No	-	-	99.7	Own	[26]
		2022	DNN	-	Supervised	1	Yes	No	100	10	94.5	Own	[27]
		2022	ML	DT	Supervised	2	No	No	-	-	99.9	Own	[28]
		2022	RNN	-	Supervised	1	No	No	-	-	97.5	Own	[29]
	IT	2009	ML	C4.5, AdaBoost, NB, SVM, RIPPER	Supervised	8	Yes	No	-	-	98.4	DARPA [15], AMP [30], MAWI [31]	[32]
		2017	CNN	1D-CNN	Supervised	12	Yes	No	40	7	99.5	ISCX VPN-nonVPN [19]	[33]
		2018	CNN	CNN-LSTM	Supervised	9	Yes	No	30	8	91.0	ISCX VPN-nonVPN [19]	[34]
		2021	Fuzzy Inference System	Fuzy Inference System	-	6	Yes	Yes	-	-	90.9	ISCXVPN2016 [33]	[35]

A "-" indicates points that were not included in the publication.

4. Traffic Analysis Methods

Protocol Classification is a term typically used to describe the process by which network traffic is classified into different categories or classes based on the characteristics of the communication. Classification can be made based on factors such as ports, addresses, packet headers, or traffic patterns. Classification aims to understand network traffic better and allow different levels of network management policy to manage this traffic as needed.

Protocol Recognition is a term usually used to describe the process by which the characteristics of network communications are analyzed to identify the protocols in use. This process can be automated using a variety of techniques, including in-depth examination of network traffic and pattern matching against a database of known protocols. The goal is to identify what protocols are used within a given communication.

Protocol Identification is a term often used as a synonym for protocol recognition, but it can also be used in a more specific sense when referring to the process of determining specific attributes or properties of a protocol that are observed in a given network traffic. Protocol identification can be important for a number of purposes, including security analysis, network optimization, and performance tuning.

4.1. Traffic Classification Technique

Several methods for traffic classification exist, each handling traffic information differently. These techniques are port-based classification, payload-based classification, statistical-based classification, behavioral-based classification, and correlation-based classification [36,37].

The port-based classification method is widely used for classifying traffic using the ports of the corresponding applications. The method is based on examining packet headers and comparing port numbers of registered applications. Examining only the packet headers presents a fast and simple classification [36]. This type of classification is especially important for identifying network applications in large network traffic [36]. The false negative rate increases because of dynamic port numbers and the use of non-standard applications. Similarly, if applications are hidden behind a commonly known port, the false positive rate increases. In general, this classification method is fast and simple, provided the applications are used with their usual ports [37].

The payload-based classification method mainly uses the packet's data content for protocol recognition. The payload information contains characteristic patterns, messages, or protocol-specific data structures [36]. Payload-based classification can be divided into Deep Packet Inspection (DPI) and Stochastic Packet Inspection (SPI) [37]. DPI works with network traffic and packet content and achieves high accuracies in traffic classification, making it a well-known technique for traffic management, attack prevention, and overall network security analysis [37,38]. SPI is a technique complementary to DPI for classifying encrypted traffic. This method works with statistical payload information to create a pattern of protocol behavior and then automatically distinguish it from other protocols. This method achieves high accuracy in classifying encrypted data. However, it is complex and computationally intensive [38]. The method represents a slight improvement over the port-based classification method but does not achieve higher accuracy in high-speed networks. The significant disadvantage of this method is network privacy. Since the method uses data inside the packet, the confidentiality of the transmitted data and network security policies are violated.

The statistical-based classification method, unlike the packet-based method and the payload-based method, does not work with information inside the packet but measures statistical traffic parameters. Based on these statistical traffic parameters, it is possible to distinguish between different types of applications [36]. These parameters include the minimum packet size, the maximum packet size, the mean packet size, and the number of packets, etc. [37]. This method is also known as the rational-based classification method [36]. The advantage of this method is that it can efficiently recognize encrypted traffic without violating privacy. The disadvantage is a large number of parameters, which may be

redundant for classification and introduce errors in training and testing machine learning models used for pattern search in large datasets [9].

The behavioral-based classification method is based on generating and analyzing host-side communication and traffic patterns. It performs classification by observing the host behavior in the network [36]. Assuming a large amount of data, this method achieves high classification accuracy [37].

The correlation-based classification method is based on creating correlations between individual data streams. Data flows are created by aggregating packets with the same attributes, such as source and destination IP address, source and destination port, and protocol used. This method is used in training and testing machine learning models to find relationships and data features. This method avoids the problem of the large number of features. However, it still poses a large computational cost [37].

4.2. Network Protocol Recognition and Identification Techniques

The basic principle in protocol recognition and identification is to extract important traffic information from the traffic, based on which the protocol can be identified. There are several methods that can be combined in protocol recognition and identification.

One division is into manual and automatic analysis. Manual analysis depends on the knowledge and experience of the person who performs the analysis [9]. Automatic analysis is based on the automatic extraction of protocol information from network traffic. Based on the extracted information, patterns of protocol behavior are created, and the techniques that enable automatic analysis create and operate on these patterns. Recognition by automatic analysis can be performed using several techniques, namely preset rules recognition, payload feature recognition, host behavior, and machine learning [9].

The preset rules recognition technique works with set rules such as port number. This method is not very reliable in terms of user customization of network settings. The payload features technique takes advantage of the deep packet inspection method, which means that it recognizes protocols using the data inside the packets. This method is very simple and easy to implement, but it cannot identify encrypted traffic and is computationally intensive. The host behavior technique works based on statistical parameters of network traffic. The method effectively avoids the process of extracting information from packets. However, the results are often inaccurate due to non-standard traffic parameters [37].

Machine learning is an important artificial intelligence technique that is used to analyze large-volume datasets based on features and associations between parameters [9]. Machine learning can be divided into shallow learning and deep learning. Shallow learning is used for modeling and analyzing. These are algorithms that cannot fully express complex nonlinear problems. At the same time, the quality of data preparation is crucial and affects the training and results of the model. Deep learning algorithms are able to solve more complex nonlinear problems. The disadvantage of classification based on shallow learning is that the feature extraction and learning process must be repeated after the dataset is changed. In deep learning-based classification, the model does not always need to be re-learned and takes advantage of the original parameters. Shallow learning algorithms include Support Vector Machine (SVM), Naive Bayes (NB), etc. Deep learning algorithms include deep neural networks, Long Short Term Memory algorithms [39], and Generative Adversarial Networks algorithms [16].

The use of deep learning methods has its application in IoT applications [40]. These methods can be used not only for anomaly detection within industrial networks but also for various operations requiring a high level of abstraction and the ability to understand complex structures.

4.3. Machine Learning Techniques for Traffic Classification

There are several types of learning: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the processed data are labeled in advance to improve the learning process and improve the final model. In teacher-less learning, data

are unlabeled and patterns are sought during the learning process by clustering the data. Feedback learning works with an agent that replaces human operators and helps determine the outcome (build a model) based on feedback [9,41].

Machine learning algorithms work with three types of data training, validation, and testing. Training and validation data form the dataset that is used to learn the model. With the help of training data, the model is created, and with the help of validation data, the model is tested during the learning process to improve the model. After the model is created/learned, the model is tested on the test dataset. Partitioning into these groups can be conducted, for example, by cross-validation. The dataset must be balanced to avoid incorrect model learning or lack of model validation [16].

Each algorithm analyzes the dataset in a different way using regression, classification, clustering, time series, association, or anomaly detection. Examples of such algorithms are linear or logistic regression, Naive Bayes algorithm, SVM, Random Forest algorithm (RF), Gradient Boosting (GB), K-Means, K-Nearest Neighbors (KNN) or Decision Tree (DT) [16].

K-means clustering is one of the popular machine-learning techniques. It belongs to the category of unsupervised learning and aims to identify unlabeled data in different clusters. The dataset and the number of clusters by which the data will be identified are essential for the proper functioning of this algorithm. Clustering consists of three parts, namely K-cluster, distance function, and new centroid. The advantage of this method is its simple implementation. The disadvantage is the sensitivity of the method to outliers [36].

K-Nearest Neighbors is a method used to determine the distance between features for classification and regression. This technique belongs to the category of supervised learning. The advantage of this method is that it is simple and suitable for problems with multiple classifications. The disadvantages of this technique are poor performance for unbalanced datasets and high computational cost.

Naive Bayes is a robust machine learning classifier for classification and belongs to the supervised category. The method is based on the Bayes Network Theorem and is used to solve complex classification and traffic identification problems. The method has many variations that use attributes for more accurate classification. The advantage of this method is high accuracy even with inaccurate data. The disadvantage is that the required attributes are independent of each other [36,37].

Support Vector Machine is another robust machine learning method. This technique is used to classify the traffic of large amounts of data. The technique is based on hyperplane separation to achieve binary classification and is classified as supervised learning. The advantage of this method is that it can solve nonlinear and high-dimensional problems. The disadvantage of this method is the high memory cost [37].

Decision Tree is a technique belonging to the supervised learning group. Decision Tree consists of a root node, several branches, and many leaves. C4.5 and ID3 machine learning classifiers are used to construct the Decision Tree. This technique is used to classify the target variable by determining the relationship and matching between attributes and creating new variables. The advantage of this method is fast classification with little computation. The disadvantage is the ease of overfitting the model when using high dimensional data because it does not correlate with these data [36].

Random Forest is a technique composed of many decision trees that fall into the same category as supervised learning. This method's advantages include a fast training phase and the fact that it is not easy to overfit due to the already mentioned large number of Decision Trees. The disadvantage is that it is unsuitable for low-dimensional and small datasets [36].

Logistic regression (LR) is a supervised learning technique. It is used for binary classification and uses general linear regression. The advantage of this method is the fast training phase and the possibility of dynamic adjustment of the classification threshold. The disadvantage of this method is easy overfitting [37].

AdaBoost is used to create a multi-classifier by integrating several weaker classifiers. By combining several classifiers and using their advantages together, the technique is able

to achieve high classification accuracy. Additionally, there is no overfitting. However, the technique is sensitive to outliers [37].

Neural networks are based on the structure of the human nervous system. The basis of a neural network is a neuron or perceptron. These basic elements are interconnected and transmit signals to each other. Neurons form networks composed of layers. Networks are made up of an input layer, inner hidden layers, and an output layer. How inputs are converted to outputs depends on the value of weights, thresholds, transformation function, and network structure. The process in question is neural network learning. Neural network learning, as with machine learning, can be conducted with a teacher (supervised) or without a teacher (unsupervised). If a neural network has multiple layers, it is referred to as a Deep Neural Network (DNN). These algorithms include Convolutional Neural Networks (CNN) [41], Recurrent Neural Networks (RNN) [29], and Artificial Neural Networks (ANN) [16].

Table 2 makes a comparison of the most well-known and some of the most used ML approaches. These are mainly supervised approaches (this is due to the nature of having to perform partitioning into known, predefined classes). Each method has defined advantages and disadvantages. ML approaches represent an effective solution when classification needs to be performed, usually with sufficient recognition capabilities (metrics). The advantage over other approaches is the relative ease of use and the equally short time required to train the model. However, the individual results are strongly influenced by the chosen task/problem and equally strongly dependent on the chosen dataset (size, purity of records, etc.). The AdaBoost approach is deliberately not shown in the table, due to the fact that it is a combination of these approaches in order to achieve the highest quality results (metrics). Neural networks represent a more sophisticated approach that can achieve more quality metrics depending on the chosen task and, in particular, the quality of the dataset, thus creating a more robust model capable of representing more challenging structures. These approaches are well suited for large data volumes and more complex problems. However, this approach requires appropriate structure and individual parameter design (especially DNN approaches) and is also a more time and computationally intensive operation. Another advantage is that NN approaches are suitable for so-called transfer learning, where model "learning" and specific data (in this case industrial protocols) are performed. This results in a more robust model.

Table 2. Comparison of ML methods for protocol recognition.

Methods	Type	Description	Advantages	Disadvantages
K-mean	Unsupervised	Identify unlabeled data in different clusters	Simple implementation	Sensitivity to outliers
K-NN	Supervised	Determine the distance between features	Simple and suitable for classification	High computational cost
NB	Supervised	Bays Network Theorem; Complex classification	High accuracy	Required attributes are independent of each other
SVM	Supervised	Hyperplane separation for binary classification	Can solve nonlinear and high-dimensional problems	High memory cost
DT	Supervised	Classify the target variable	Fast classification	Ease of overfit
RF	Supervised	Algorithm composed of many decision trees	Fast training phase; Not easy to overfit	Unsuitable for low-dimensional and small datasets
LR	Supervised	General linear regression	Fast training phase; Not easy to overfit	Easy to overfit

4.4. Metrics for Machine Learning Model Evaluation

In order to evaluate machine learning models, it is necessary to use evaluation metrics [37,42]. These metrics numerically express the model's ability to perform defined activities, such as classification. The most basic case is binary classification, where a mapping of an input to just two outputs (0 or 1) is performed. If 1 is marked as a positive outcome (for example, a classified attack), four situations can occur:

- True Positive (TP)—(1 = 1)—the input is an attack, and the output of the model is classified as an attack,
- True Negative (TN)—(0 = 0)—the input is regular traffic, and the model output is classified as regular traffic,
- False Positive (FP)—($0 \neq 1$)—the input is regular traffic, and the output of the model is classified as an attack,
- False Negative (FN)—($1 \neq 0$)—the input is an attack, and the output of the model is classified as regular traffic.

These metrics are further used to calculate other auxiliary evaluation metrics [42]. Visualizations of the underlying metrics can be made in the form of a Confusion Matrix, which is used to provide a basic representation of the ratio of each group. Additional metrics are then typically calculated based on this matrix. Accuracy is a metric that defines the comprehensive success rate of the model and is defined as the ratio of TP and TN to the total number of classified entries; see Equation (1). This metric can be described as a definition of how good a model is and its recognition capabilities. Precision is a metric that is defined as the ratio of TP to the sum of TP and FP; see Equation (2). This metric is described as the accuracy of the model, i.e., whether the recognition capabilities are correct and whether it produces coherent results. Recall is a metric also referred to as True Positive Rate or Sensitivity and is the ratio of TP to the sum of TP and FN; see Equation (3). It is an indicator of the completeness of the model's detection of positive cases. Precision focuses on the accuracy of positive predictions, while Recall evaluates the ability of the model to detect as many true positive cases (TP) as possible. The F1 score metric provides a composite view of how accurate the model is, not only in its accuracy but also in its ability to identify TPs using the Precision and Recall metrics, see Equation (4). This metric aids model assessment, especially in cases where the dataset is unbalanced and where separate Precision and Recall values could be misleading.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}\ [-], \tag{1}$$

$$Precision = \frac{TP}{TP + FP}\ [-], \tag{2}$$

$$TPR; Sensitivity; Recall = \frac{TP}{TP + FN}\ [-], \tag{3}$$

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}\ [-]. \tag{4}$$

Another metric used is the False Positive Rate (FPR) $\frac{FP}{FP+TN}$. This metric defines the false positive rate (FP) to all actually negative cases, how the model misinterprets negative cases as positives. False Negative Rate (FNR) $\frac{FN}{FN+TP}$. Metric defines the rate of false negative cases to all actually positive cases, how the model misinterprets positive cases as negatives. The True Negative Rate or also Specificity (TNR) metric $\frac{TN}{TN+FP}$ gives the ratio of actual negative cases to all true negative cases, and how well the model can detect negative situations or events.

The TPR and FPR indicators are further used to represent graphically in the form of a Receiver Operating Characteristic (ROC) curve of the model capabilities. FPR is plotted on the X-axis, and TPR is plotted on the Y-axis. The objective is to plot the threshold values to find a compromise between the high Sensitivity (TPR) and low FPR. The resulting Area Under the Curve (AUC) allows for the assessment of the model performance (a larger AUC implies a better AI model).

As the number of recognized classes increases (input data are classified into more groups, for example, identifying a specific type of attack), it is possible to approach accuracy from different perspectives [37]. Thus, it is possible to obtain an overall accuracy, which indicates the general ability of the classification model regardless of the specific class (number/measure of appropriately labeled samples regardless of the class). Class accuracy

indicates the accuracy achieved relative to a specific class (some classes may achieve higher accuracy than others—this can identify strong/weak points of the model, hence the dataset). It is also possible to relate accuracy to the data flow itself (correlation-based classification methods) or byte accuracy. Byte accuracy focuses on individual bytes (even within a flow) [37].

5. Industrial Datasets Analysis

Datasets can be used for research in machine learning and neural networks. These datasets are used to train artificial intelligence tools and allow for the evaluation of different processing approaches [41].

This is possible just with a public dataset because it allows us to compare different approaches on an identical dataset. Thus, it is possible to identify suitable approaches such as the algorithm itself, its settings (hyperparameters), the preprocessing, or the representation of the dataset within AI processing.

AI development can also be performed on a custom dataset, but there is a risk of dataset deficiencies such as inconsistency (data may be recorded in an inappropriate way), incompleteness/diversity (not all possible states are included), duplication (dataset contains duplicate records), imbalance (representation of individual classes is not even), size (dataset is too small). Due to these problems/deficiencies, the developed AI algorithm can paradoxically achieve high portability, but in the case of practical use, such a tool is very limited (or overtraining may occur).

For this reason, datasets are published providing identical data on which the different approaches can be evaluated. However, published datasets often run into privacy issues. Thus, it is necessary to check the individual data within the dataset and ensure consent, anonymize the data, or generate the dataset in a closed/protected environment where sensitive information cannot be leaked.

Datasets can be stored in various data formats, the most common of which is the Comma-Separated Values (CSV) format or the network traffic record—PCAP format. Where the CSV format is more strict in terms of available information (features), the PCAP format allows parsing a wide range of information. Another important aspect of datasets is the documentation available. The documentation should include a complete description of the dataset, including a description of the main components (IP addresses, transport ports, etc.), in particular, the number of classes and the way the dataset is labeled, especially in the case of the CSV format. In the case of the PCAP format, it is important to uniquely identify the individual states that occurred in the record (especially in cyber-security)—e.g., identify the attacker, their IP address, ports, time horizon, type of attack, etc. Other useful data are the wiring/schema used to generate the dataset, tools used, etc.

Table 3 performs a comparison of publicly available datasets and makes comparisons from several perspectives. A comparison of the IT/OT focus of the dataset is made; the number of classes into which classification is made, the number of features (for CSV format only), and the format of the dataset is given. It also found whether the dataset is time-series data, whether the dataset is labeled, and the type of classification (anomalies—cyber-security attacks, protocol classification, OT anomalies). In addition, whether the dataset is cyber-security focused, the source is indicated, whether it is a real record or a simulation, and the protocol (IT represents common IT protocols). Last but not least, the number of records (CSV only) and whether documentation is available.

A total of 28 datasets were compared, where 17 datasets fall into the IT sector and 11 into the OT sector. Sixteen datasets are recorded as PCAP and 17 as CSV. Similarly, 17 datasets focus on time-series. The analysis also shows that a large part of the datasets focuses on the problem and anomalies as well as binary classification. The datasets directly targeting the problem of protocol analysis and classification form a very small part, with three datasets out of 17 in IT and only one dataset out of 11 in OT. The bulk is also not recorded directly in the real environment, which is due to the general focus on anomaly identification (or classification), but the recording is as close as possible to the real environ-

ment using simulated parts. All datasets contain documentation but often do not contain all the necessary information.

The analysis shows that the majority of available public datasets are focused on cybersecurity issues, while the classification of protocols is minimal. This may also be due to the distinct individuality of individual plants and industries. To this end, it is thus nontrivial to use the available datasets and to use only normal/regular data flow, thus reducing the dataset size considerably. Another aspect is the availability of public datasets, where some datasets are difficult to access, and it is necessary to be a subscriber.

In addition, datasets are available from the OT environment that contain only sensory data. Thus, these datasets cannot be directly used for the detection of industrial hazards but only for the detection or classification of security incidents within a given workplace. These datasets may include, among others, those mentioned in [43]. However, signal data cannot be used for protocol classification purposes, and anomaly detection may be strongly associated with a given workplace. It is necessary to train the AI model on just the specific states that can be "accepted"/assumed in a given environment.

Thus, based on the analysis performed, individual challenges were identified in order to classify the industrial protocols:

(i) create a representative and comprehensive dataset using an industrial protocol,
(ii) to use real industrial networks and devices to get closer to real applications,
(iii) to allow modification or change of the industrial protocol (within the dataset)—protocol diversity.

Fulfilling these challenges will thus enable research into methods to classify industrial protocols, the protocol versions used, and the cipher suites of encrypted protocol versions used. In the case of using the same workstation with only a change of industrial protocol, it is possible to focus only on the protocol itself, without the influence of the transmitted data on the industrial protocol classification performed (the protocol classification will not be directly influenced by the transmitted data).

A total of 28 datasets were compared in the analysis, with most of them focusing only on a specific part and, in particular, on areas of cyber security anomalies. The classification of the protocols themselves is very limited, especially in the OT area. In the case of the use of IT protocols, typical IT protocols such as HTTP, HTTPS, DNS, FTP, ICMP, IMAP, POP3, etc., are often used. In the case of OT protocols, these are typically Modbus, IEC 60870-5-104, and DNP3. There is also a range of sensor data (data obtained from the L0 Purdue model—data obtained from sensors and actuators). However, these data are intended for anomaly detection in terms of the behavior of individual states and do not contain the protocol itself (they are only application data or values of individual variables). In terms of size (records, volume parameter in Table 3), the individual datasets vary considerably, and in the case of the PCAP source, this value is variable depending on what data are parsed. Dataset balance has not been considered in the table because many datasets do not provide predefined training and test sets (separate datasets). It is the train/test split parameter that may be crucial and, as such, it may be a target of research.

Table 3. Overview of the most relevant datasets for machine learning and neural network research.

Link	Name of Dataset	Year	IT/OT	Classes	Feature Count	Format	Time-Series	Labeled	Classification of	Cyber-sec.	Source	Protocol	Volume	Docu.
[15]	DARPA	1998	IT	2	NR	PCAP	Yes	No	Anomaly	Yes	Real *	IT	-	Yes
[44]	KDD Cup 1999	1999	IT	5	41	CSV	No	Yes	Anomaly	Yes	Simulated	-	4,000,000	Yes *
[31]	MAWI/Wide/Keio	2000	IT	?	NR	PCAP	Yes	Yes *	Protocol	No	Real *	IT	-	Yes *
[45]	CAIDA	2008	IT	?	NR	PCAP	Yes	No	Protocol	No	Real	IT	-	Yes *
[46]	NSL-KDD	2009	IT	2	41	CSV	No	Yes	Anomaly	Yes	Simulated	-	148,000	Yes
[47]	MAWILab	2010	IT	4	NR	PCAP	Yes	Yes	Anomaly	Yes	Real *	IT	-	Yes
[48]	ISCX-IDS-2012	2012	IT	2	NR	PCAP	Yes	Yes	Anomaly	Yes	Real *	IT	-	Yes
[49]	CTU-13	2014	IT	3	NR	PCAP; BIGARUS	Yes	Yes	Anomaly	Yes	Real *	IT	-	Yes
[50]	ISCX-Bot-2014	2014	IT	2	NR	PCAP	Yes	Yes	Anomaly	Yes	Real *	IT	-	Yes
[51]	UNSW-NB15	2015	IT	10	49	CSV	No	Yes	Anomaly	Yes	Real *	IT	2,500,000	Yes
[52]	CTU-Mixed (capture 1–8)	2015	IT	2	NR	PCAP; BIGARUS	Yes	No	Anomaly	Yes	Real	IT	-	Yes
[53]	USTC-TFC2016	2016	IT	20	NR	PCAP	Yes	Yes	Protocol; Anomaly	Yes	Real *	IT	-	Yes
[54]	CIC-IDS-2017	2017	IT	2	78	CSV	No	Yes	Anomaly	Yes	Real *	IT	692,703	Yes
[55]	CAN 2017	2017	IT	4	11	TXT	No	Yes *	OT anomaly	Yes	Real *	CAN	4,613,909	Yes
[54]	CSE-CIC-IDS2018	2018	IT	7	80	CSV	No	Yes	Anomaly	Yes	Real *	IT	16,233,002	Yes
[56]	CIRA-CIC-DoHBrw-2020	2020	IT	2	34	CSV	No	Yes	Anomaly	Yes	Real *	IT	371,836	Yes
[57]	NSS Mirai	2021	IT	11	12	CSV	No	Yes	Anomaly	Yes	Real *	IT	64,025	Yes *
[58]	Electra dataset	2010	OT	4	10	CSV	No	Yes	OT anomaly	Yes	Simulated	Modbus, S7comm	1,048,575	Yes *
[23]	SWAT	2015	OT	2	NR	PCAP; CSV	Yes	Yes	OT anomaly	Yes	Real	Senzoric data	-	Yes *
[19]	ISCX VPN-nonVPN	2016	IT/OT	14	NR	PCAP; CSV	Yes	Yes	Protocol	No	Real *	IT	-	Yes
[59]	Batadal	2016	OT	2	45	CSV	Yes	Yes *	OT anomaly	Yes	Real *	Senzoric data	23,788	Yes *
[24]	Providing SCADA Network Data Sets for Intrusion Detection Research	2016	OT	2	NR	PCAP; CSV	Yes	Yes *	OT anomaly	Yes	Real *	Modbus; Senzoric data	-	Yes *
[60]	WADI	2017	OT	2	NR	PCAP; CSV	Yes	Yes	OT anomaly	Yes	Real	Senzoric data	1,221,372	Yes *
[61]	BoT-IoT	2019	OT	5	46	CSV	No	Yes	Anomaly	Yes	Real *	IT	72,000,000	Yes
[62]	DNP3 Intrusion Detection Dataset	2022	OT	?	NR	PCAP; CSV	Yes	Yes	OT anomaly	Yes	?	DNP3	-	Yes *
[63]	CIC Modbus dataset 2023	2023	OT	?	NR	PCAP	Yes	No	OT anomaly	Yes	Simulated	Modbus	-	Yes
[64]	IEC 60870-5-104 Intrusion Detection Dataset	2023	OT	?	NR	PCAP; CSV	Yes	Yes	OT anomaly	Yes	?	IEC 60870-5-104	-	Yes *
[65]	HIL-based augmented ICS security	2023	OT	53	225 (HAIEnd)	CSV	No	Yes	OT anomaly	Yes	Real	Senzoric data	?	Yes *

* indicates incomplete fulfilment of the criterion; NR = Not Relevant; ? = unable to find.

6. Discussion

The purpose of this paper was to answer the two main scientific questions presented in the introduction. How can industrial protocol classification be achieved? What publicly available datasets can currently be used specifically to classify these protocols? Classification, recognition, and identification of protocols are closely related techniques that use the same methods. The most commonly used methods for traffic classification and protocol recognition have been presented and compared. Each method has its specific use and depends on the purpose for which it is to be used. Among the state-of-the-art methods are machine learning algorithms and especially neural networks. These algorithms allow for fast traffic classification and protocol recognition and provide high-quality metrics. However, these algorithms are limited in terms of input data. An analysis of the state of the art revealed that the majority of research is in the IT domain. Similarly, research is not targeted at encrypted versions of protocols.

The available datasets often do not achieve the qualities needed for good and accurate classification, such as the number of records, the diversity of records, or the number of logs in the dataset. Currently, the number of datasets from IT environments exceeds the number of datasets. Although it is possible to use some IT protocols in OT systems from the point of view of the convergence of IT and OT networks, it is not advisable to rely on this fact alone. OT networks require specific protocols and requirements that are not as strict in IT networks. Based on the analysis of publicly available datasets, key requirements for future research were identified. Namely, the creation of a representative dataset containing industrial protocols using real industrial devices. Currently, no suitable dataset has been found for protocol recognition research in OT. It is the creation of such a dataset that would enable follow-up research and the comparison of different methods from the ML and NN domains.

7. Conclusions

The issue of protocol recognition and traffic classification is a broad area with overlap from IT to OT networks. In conjunction with the convergence of IT and OT networks, it is necessary to focus on cyber-security within OT networks and to use current techniques from IT and implement them in the OT domain in order to increase the current level of security. Similarly, with the trend of Industry 4.0+, data (not only IT but also OT) are leaving isolated networks for processing on remote servers or for using software as a service. For this reason, this paper has focused on the analysis of different methods and processing of data flow (or other units) for the purpose of protocol recognition and traffic classification in connection with OT specifics. Furthermore, publicly available datasets have been compared in terms of their contribution, usability, etc. The output of this work is thus a comparative analysis of approaches specifically to protocol recognition and traffic classification. The analysis shows that there is currently only a very limited number of publicly available datasets that would allow development in the area of protocol recognition and traffic classification in OT networks. Thus, it is necessary to build on the IT networks and the knowledge gained in the area of protocol recognition and traffic classification in IT networks and, on the basis of a good and robust dataset, to compare these approaches, to make modifications and, in particular, to evaluate them in OT networks.

Author Contributions: Conceptualization, E.H., R.F. and J.M.; methodology, E.H. and R.F.; validation, E.H., R.F. and J.M.; formal analysis, E.H. and R.F.; investigation, E.H.; resources, E.H.; data curation, E.H. and R.F.; writing—original draft preparation, E.H.; writing—review and editing, E.H., R.F. and J.M.; visualization, E.H. and R.F.; supervision, R.F. and J.M.; project administration, R.F. and J.M.; funding acquisition, R.F. All authors have read and agreed to the published version of the manuscript.

Funding: This article is a result of the project FW07010004, which was supported by the Technology Agency of the Czech Republic in the Program TREND.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Santos, M.F.O.; Melo, W.S.; Machado, R. Cyber-Physical Risks identification on Industry 4.0. In Proceedings of the 2022 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT), Trento, Italy, 7–9 June 2022; pp. 300–305. [CrossRef]
2. Santos, S.; Costa, P.; Rocha, A. IT/OT Convergence in Industry 4.0. In Proceedings of the 2023 18th Iberian Conference on Information Systems and Technologies (CISTI), Aveiro, Portugal, 20–23 June 2023; pp. 1–6. [CrossRef]
3. Duan, L.; Da Xu, L. Data Analytics in Industry 4.0: A Survey. *Inf. Syst. Front.* 2021, *ahead of print*. [CrossRef]
4. Knapp, E.D.; Langill, J.T. Chapter 8—Risk and Vulnerability Assessments. In *Industrial Network Security*, 2nd ed.; Knapp, E.D., Langill, J.T., Eds.; Syngress: Boston, MA, USA, 2015; pp. 1–439.
5. Parsons, D. *SANS ICS/OT Cybersecurity Survey: 2023's Challenges and Tomorrow's Defenses, Sans.org*; SANS Institute: Rockville Pike, MD, USA, 2023; pp. 1–19.
6. *ISA-99—Industrial Automation and Control Systems Security*; International Society of Automation (ISA): Pittsburgh, PA, USA, 2007.
7. Perducat, C.; Mazur, D.C.; Mukai, W.; Sandler, S.N.; Anthony, M.J.; Mills, J.A. Evolution and Trends of Cloud on Industrial OT Networks. *IEEE Open J. Ind. Appl.* **2023**, *4*, 291–303. [CrossRef]
8. Grüner, S.; Trosten, A. A Cloud-Native Software Architecture of NAMUR Open Architecture Verification of Request using OPC UA PubSub Actions over MQTT. In Proceedings of the 2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA), Sinaia, Romania, 12–15 September 2023; pp. 1–8. [CrossRef]
9. Zhai, L.; Zheng, Q.; Zhang, X.; Hu, H.; Yin, W.; Zeng, Y.; Wu, T. Identification of Private ICS Protocols Based on Raw Traffic. *Symmetry* **2021**, *13*, 1743. [CrossRef]
10. Ning, B.; Zong, X.; He, K.; Lian, L. PREIUD: An Industrial Control Protocols Reverse Engineering Tool Based on Unsupervised Learning and Deep Neural Network Methods. *Symmetry* **2023**, *15*, 706. [CrossRef]
11. Chen, C.; Wang, F.; Lin, F.; Guo, S.; Gong, B. Fast Protocol Recognition by Network Packet Inspection. *Neural Inf. Process.* **2011**, *7063*, 37–44. [CrossRef]
12. Liu, Q.; Zhang, J.; Zhao, B. Traffic Classification Using Compact Protocol Fingerprint. In Proceedings of the 2012 International Conference on Industrial Control and Electronics Engineering, Xi'an, China, 23–25 August 2012; pp. 147–151. [CrossRef]
13. Vulnerability Databases. *Rediris.es* 2001. Available online: https://www.rediris.es/cert/links/vuldb.html.en (accessed on 21 March 2024).
14. Lopez-Martin, M.; Carro, B.; Sanchez-Esguevillas, A.; Lloret, J. Network Traffic Classifier With Convolutional and Recurrent Neural Networks for Internet of Things. *IEEE Access* **2017**, *5*, 18042–18050. [CrossRef]
15. Lippmann, R.; Haines, J.W.; Fried, D.J.; Korba, J.; Das, K. Analysis and Results of the 1999 DARPA Off-Line Intrusion Detection Evaluation. *Recent Adv. Intrusion Detect.* **2000**, *1907*, 162–182. [CrossRef]
16. Feng, W.; Hong, Z.; Wu, L.; Fu, M.; Li, Y.; Lin, P. Network protocol recognition based on convolutional neural network. *China Commun.* **2020**, *17*, 125–139. [CrossRef]
17. Xue, J.; Chen, Y.; Li, O.; Li, F. Classification and identification of unknown network protocols based on CNN and T-SNE. *J. Phys. Conf. Ser.* **2020**, *1617*, 012071. [CrossRef]
18. Shi, J.; Yu, X.; Liu, Z.; Niu, B. Nowhere to Hide. *Secur. Commun. Netw.* **2021**, *2021*, 6672911. [CrossRef]
19. Draper-Gil, G.; Lashkari, A.H.; Mamun, M.S.I.; Ghorbani, A.A. Characterization of Encrypted and VPN Traffic using Time-related Features. In Proceedings of the 2nd International Conference on Information Systems Security and Privacy, Rome, Italy, 19–21 February 2016; pp. 407–414. [CrossRef]
20. Lu, B.; Luktarhan, N.; Ding, C.; Zhang, W. ICLSTM. *Symmetry* **2021**, *13*, 1080. [CrossRef]
21. Zhu, P.; Wang, G.; He, J.; Chang, Y.; Kong, L.; Liu, J. Encrypted Traffic Protocol Identification Based on Temporal and Spatial Features. In Proceedings of the 2023 4th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Nanjing, China, 16–18 June 2023; pp. 255–262. [CrossRef]
22. de Toledo, T.; Torrisi, N. Encrypted DNP3 Traffic Classification Using Supervised Machine Learning Algorithms. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 384–399. [CrossRef]
23. Mathur, A.P.; Tippenhauer, N.O. SWaT. In Proceedings of the 2016 International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWater), Vienna, Austria, 11 April 2016; pp. 31–36. [CrossRef]
24. Lemay, A.; Fernandez, J.M. Providing SCADA Network Data Sets for Intrusion Detection Research. In Proceedings of the 9th Workshop on Cyber Security Experimentation and Test (CSET 16), Austin, TX, USA, 8 August 2016.
25. Sheng, C.; Yao, Y.; Yang, W.; Liu, Y.; Fu, Q. How to Fingerprint Attack Traffic against Industrial Control System Network. In Proceedings of the 2019 1st International Conference on Industrial Artificial Intelligence (IAI), Shenyang, China, 23–27 July 2019; pp. 1–6. [CrossRef]
26. Lan, H.; Zhu, X.; Sun, J.; Li, S. Traffic Data Classification to Detect Man-in-the-Middle Attacks in Industrial Control System. In Proceedings of the 2019 6th International Conference on Dependable Systems and Their Applications (DSA), Harbin, China, 23–27 July 2020; pp. 430–434. [CrossRef]

27. Holasova, E.; Fujdiak, R. Deep Neural Networks for Industrial Protocol Recognition and Cipher Suite Used. In Proceedings of the 2022 IEEE International Carnahan Conference on Security Technology (ICCST), Valec, Czech Republic, 7–9 September 2022; pp. 1–7. [CrossRef]
28. Yu, C.; Zhang, Z.; Gao, M. An ICS Traffic Classification Based on Industrial Control Protocol Keyword Feature Extraction Algorithm. *Appl. Sci.* **2022**, *12*, 11193. [CrossRef]
29. Wang, W.; Zhang, B.; Yu, Z.; Gao, X. Anomaly Detection Method of Unknown Protocol in Power Industrial Control System Based on RNN. In Proceedings of the 2022 5th International Conference on Renewable Energy and Power Engineering (REPE), Beijing, China, 28–30 September 2022; pp. 68–72. [CrossRef]
30. Zhang, F.; Wei, K.; Slowikowski, K.; Fonseka, C.Y.; Rao, D.A.; Kelly, S.; Goodman, S.M.; Tabechian, D.; Hughes, L.B.; Salomon-Escoto, K.; et al. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nat. Immunol.* **2019**, *20*, 928–942. [CrossRef] [PubMed]
31. Cho, K. MAWI Working Group Traffic Archive. Available online: http://mawi.wide.ad.jp/mawi/ (accessed on 20 March 2024).
32. Alshammari, R.; Zincir-Heywood, A.N. Machine learning based encrypted traffic classification. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 8–10 July 2009; pp. 1–8. [CrossRef]
33. Wang, W.; Zhu, M.; Wang, J.; Zeng, X.; Yang, Z. End-to-end encrypted traffic classification with one-dimensional convolution neural networks. In Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China, 22–24 July 2017; pp. 43–48. [CrossRef]
34. Zou, Z.; Ge, J.; Zheng, H.; Wu, Y.; Han, C.; Yao, Z. Encrypted Traffic Classification with a Convolutional Long Short-Term Memory Neural Network. In Proceedings of the 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Exeter, UK, 28–30 June 2018; pp. 329–334. [CrossRef]
35. Kim, S.W.; Kim, K.C. Traffic Type Recognition Method for Unknown Protocol—Applying Fuzzy Inference. *Electronics* **2021**, *10*, 36. [CrossRef]
36. Sheikh, M.S.; Peng, Y. Procedures, Criteria, and Machine Learning Techniques for Network Traffic Classification: A Survey. *IEEE Access* **2022**, *10*, 61135–61158. [CrossRef]
37. Zhao, J.; Jing, X.; Yan, Z.; Pedrycz, W. Network traffic classification for data fusion. *Inf. Fusion* **2021**, *72*, 22–47. [CrossRef]
38. Xu, C.; Chen, S.; Su, J.; Yiu, S.M.; Hui, L.C.K. A Survey on Regular Expression Matching for Deep Packet Inspection: Applications, Algorithms, and Hardware Platforms. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 2991–3029. [CrossRef]
39. Zhao, H.; Li, Z.; Wei, H.; Shi, J.; Huang, Y. SeqFuzzer: An Industrial Protocol Fuzzing Framework from a Deep Learning Perspective. In Proceedings of the 2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST), Xi'an, China, 22–27 April 2019; pp. 59–67. [CrossRef]
40. Elhanashi, A.; Dini, P.; Saponara, S.; Zheng, Q. Integration of Deep Learning into the IoT. *Electronics* **2023**, *12*, 4952. [CrossRef]
41. Krupski, J.; Graniszewski, W.; Iwanowski, M. Data Transformation Schemes for CNN-Based Network Traffic Analysis: A Survey. *Electronics* **2021**, *10*, 2042. [CrossRef]
42. Yan, J. A Survey of Traffic Classification Validation and Ground Truth Collection. In Proceedings of the 2018 8th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 15–17 June 2018; pp. 255–259. [CrossRef]
43. Jourdan, N.; Longard, L.; Biegel, T.; Metternich, J. Machine Learning for Intelligent Maintenance and Quality Control: A Review of Existing Datasets and Corresponding Use Cases. In Proceedings of the Conference on Production Systems and Logistics: CPSL 2021, Hannover, Germany, 25–28 May 2021; Volume 2. [CrossRef]
44. Salvatore, S.; Wei, F.; Wenke, L.; Andreas, P.; Philip, C. *KDD Cup 1999 Data*; UCI Machine Learning Repository: Irvine, CA, USA 1999. [CrossRef]
45. UCSD C. *The CAIDA Anonymized Internet Traces Dataset (April 2008–January 2019)*; CAIDA: La Jolla, CA, USA, 2018.
46. Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A.A. A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 8–10 July 2009; pp. 1–6. [CrossRef]
47. Fontugne, R.; Borgnat, P.; Abry, P.; Fukuda, K. MAWILab. In Proceedings of the 6th International COnference, New York, NY, USA, 26–28 August 2010; pp. 1–12. [CrossRef]
48. Shiravi, A.; Shiravi, H.; Tavallaee, M.; Ghorbani, A.A. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Comput. Secur.* **2012**, *31*, 357–374. [CrossRef]
49. García, S.; Grill, M.; Stiborek, J.; Zunino, A. An empirical comparison of botnet detection methods. *Comput. Secur. J.* **2014**, *45*, 100–123. [CrossRef]
50. Beigi, E.B.; Jazi, H.H.; Stakhanova, N.; Ghorbani, A.A. Towards effective feature selection in machine learning-based botnet detection approaches. In Proceedings of the 2014 IEEE Conference on Communications and Network Security, San Francisco, CA, USA, 29–31 October 2014; pp. 247–255. [CrossRef]
51. Moustafa, N.; Slay, J. UNSW-NB15. In Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, 10–12 November 2015; pp. 1–6. [CrossRef]
52. Garcia, S. Malware Capture Facility Project, 2018. Available online: https://stratosphereips.org (accessed on 20 March 2024).

53. Wang, W.; Zhu, M.; Zeng, X.; Ye, X.; Sheng, Y. Malware traffic classification using convolutional neural network for representation learning. In Proceedings of the 2017 International Conference on Information Networking (ICOIN), Da Nang, Vietnam, 11–13 January 2017; pp. 712–717. [CrossRef]
54. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy, Funchal, Portugal, 22–24 January 2018; pp. 108–116. [CrossRef]
55. Lee, H.; Jeong, S.H.; Kim, H.K. OTIDS. In Proceedings of the 2017 15th Annual Conference on Privacy, Security and Trust (PST), Calgary, AB, Canada, 28–30 August 2017; pp. 57–5709. [CrossRef]
56. MontazeriShatoori, M.; Davidson, L.; Kaur, G.; Lashkari, A.H. Detection of DoH Tunnels using Time-series Classification of Encrypted Traffic. In Proceedings of the 2020 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Calgary, AB, Canada, 17–22 August 2020; pp. 63–70. [CrossRef]
57. Kalupahana Liyanage, K.S.; Divakaran, D.M.; Singh, R.P.; Gurusamy, M. NSS Mirai Dataset . Available online: https://ieee-dataport.org/documents/nss-mirai-dataset (accessed on 20 March 2024).
58. Electra Dataset: Anomaly Detection ICS Dataset. Available online: http://perception.inf.um.es/ICS-datasets/ (accessed on 20 March 2024).
59. Taormina, R.; Galelli, S.; Tippenhauer, N.O.; Salomons, E.; Ostfeld, A.; Eliades, D.G.; Aghashahi, M.; Sundararajan, R.; Pourahmadi, M.; Banks, M.K.; et al. Battle of the Attack Detection Algorithms. *J. Water Resour. Plan. Manag.* **2018**, *144*, 1–11 . [CrossRef]
60. Ahmed, C.M.; Palleti, V.R.; Mathur, A.P. WADI. In Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks, Pittsburgh, PA, USA, 21 April 2017; pp. 25–28. [CrossRef]
61. Koroniotis, N.; Moustafa, N.; Sitnikova, E.; Turnbull, B.P. Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset. *arXiv* **2018**, arXiv:1811.00701.
62. Radoglou-Grammatikis, P.; Kelli, V.; Lagkas, T.; Argyriou, V.; Sarigiannidis, P. DNP3 Intrusion Detection Dataset. 2022. Available online: https://ieee-dataport.org/documents/dnp3-intrusion-detection-dataset (accessed on 20 March 2024).
63. Boakye-Boateng, K.; Ghorbani, A.A.; Lashkari, A.H. Securing Substations with Trust, Risk Posture, and Multi-Agent Systems. In Proceedings of the 2023 20th Annual International Conference on Privacy, Security and Trust (PST), Copenhagen, Denmark, 21–23 August 2023; pp. 1–12. [CrossRef]
64. Radoglou-Grammatikis, P.; Rompolos, K.; Lagkas, T.; Argyriou, V.; Sarigiannidis, P. IEC 60870-5-104 Intrusion Detection Dataset. 2022. Available online: https://ieee-dataport.org/documents/iec-60870-5-104-intrusion-detection-dataset (accessed on 20 March 2024).
65. Shin, H.K.; Lee, W.; Yun, J.H.; Kim, H. HAI 1.0: HIL-based Augmented ICS Security Dataset. In Proceedings of the 13th USENIX Workshop on Cyber Security Experimentation and Test (CSET 20), Online, 10 August 2020; USENIX Association: Berkeley, CA, USA, 2020.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.