Edge AI Assurance: A Systematic Mapping Study

Clara Ayora[1*], Arturo S. García[1], and Jose Luis de la Vara[1]

[1]*Universidad de Castilla-La Mancha, Avda. España s/n, 02071 Albacete, (Spain)*

Context. In critical domains, assurance corresponds to the set of activities to provide confidence that a system can be deemed dependable, e.g., safe and secure. This essential system and software engineering process is usually conducted according to standards. For novel applications running at the edge and using artificial intelligence (AI), how to conduct assurance in a systematic way is still under study.

Objective. The goal of this paper is to provide a comprehensive understanding of current Edge AI assurance considerations. Our interest lies in contributing insights that offer a forward-looking perspective on what is essential in this research field.

Method. We conducted a systematic mapping study (SMS) to characterize how Edge AI assurance is addressed in existing literature. The SMS was built on 38 studies, selected through a multi-stage process, from 3113 studies published between 2019 and 2025. The 38 studies were deeply analysed focusing on seven research questions about the main key Edge AI assurance aspects: dependability concerns, application domains, assurance standards, assurance evidence, dependability justification techniques, and edge and AI characteristics.

Results. We found ten dependability concerns that have been addressed (e.g., safety and security), six application domains (e.g., Industry 4.0), eight assurance standards and regulations (e.g., ISO 26262), 27 types of assurance evidence (e.g., architecture specification), three dependability justification techniques (e.g., argumentation), five AI-specific characteristics (e.g., machine learning algorithms) and five edge-specific characteristics (e.g., network).

Conclusions. The paper is, to our knowledge, the only existing review on the topic of Edge AI assurance. The results are relevant to practitioners seeking a better grasp on this field as well as researchers to find new research gaps. We have also identified research areas where more effort can be undertaken (e.g., multi-concern assurance).

Keywords: Edge AI, Assurance, Systematic Mapping Study.

1. Introduction

Among the current and future key Information and Communication Technologies [22], edge computing has emerged as an architectural paradigm that brings computation and data storage closer to

* Corresponding author
Email address: clara.ayora@uclm.es (Clara Ayora)

data sources and that is expected to save time and bandwidth. Another key technology is artificial intelligence (AI), which is facilitating and will continue to facilitate the automation of processes that are largely reliant on human cognitive abilities and the execution of complex cognitive tasks that humans are unable to perform. The use of AI on the edge can be referred to as Edge AI [69]. Edge AI applications are gaining attention nowadays and are starting to be developed and deployed for a wide range of contexts and application domains, including critical ones, e.g., automotive, aerospace, healthcare, and Industry 4.0. When Edge AI applications, as well as others, perform critical functions (whose failure may negatively impact on e.g., safety or security, or other dependability attributes), they are subject to system assurance [50].

Assurance can be defined as the set of activities to provide adequately justified confidence that a system satisfies given requirements, e.g., for system safety and security, thus for system dependability [20]. For instance, system assurance is needed to have confidence that an autonomous system will not harm someone. This confidence is often developed by satisfying certain objectives that mitigate the potential risks that a system can pose during its lifecycle. This is usually performed in compliance with assurance and engineering standards, e.g., IEC 61508 [32]. However, for systems including AI at the edge, and although several initiatives aim to tackle Edge AI complexity (e.g., AI models may change and deteriorate over time) and to ensure its quality, a unified standard remains elusive [30].

For cost-effective assurance of novel critical Edge AI applications, several issues need to be addressed. Firstly, it is not sufficient to address only one concern [26], such as safety, but security must also be considered due to the connectivity of edge devices. Secondly, specific Edge AI characteristics must be taken into account [67], including where and how data is generated, stored, and used, where and how decisions are made, and the dedicated software and hardware developed, such as accelerators and hypervisors. Finally, the assurance needs of novel applications must be addressed [8], considering the distinctive characteristics of Edge AI applications. For example, the assurance requirements of a drone monitoring infrastructure may differ from those of a system that interacts with people. In general, the assurance of new technologies usually poses challenges in practice due to the lack of defined and agreed-upon best practices that address their specific and distinct characteristics [55].

The main objective of this paper is to synthesise the existing knowledge in the literature about Edge AI assurance. This is performed by means of a systematic mapping study (SMS) – a research method used to systematically review and categorize existing literature within a specific field [47,48,62]. The main advantage of a SMS, when compared to ad hoc search, is that it provides a higher degree of confidence about covering the relevant literature and thus minimises subjectivity and bias. Without this synthesis (i.e., the SMS), advancing in research could become more challenging, leading to redundancies in research efforts and hindering the effective resolution of emerging challenges in Edge AI.

The key contributions of this paper include:

- A comprehensive overview of the current state of research about Edge AI assurance. This includes studying the addressed dependability concerns (e.g., safety, security, and trustworthiness), standards that can be used for Edge AI assurance, types of evidence and justification techniques used for such assurance, and the specific aspects of Edge AI that have been considered.

- A summary of findings and limitations identified from the studies, including research gaps to guide future studies.

- Support for researchers and practitioners in gaining comprehensive understanding of Edge AI assurance.

The SMS has been conducted as part of REBECCA (https://rebecca-chip.eu/), a large-scale EU project on a new reconfigurable, heterogeneous and highly parallel processing platform for safe and secure AI. Assurance is a part of the overall work of the project to enable efficient Edge AI solutions that can overcome physical limitations, are dependable, and enhance European strategic autonomy. As a whole, REBECCA considers hardware and software development for Edge AI, their integration and validation, their safety and security, their use in different systems, and compliance. In a prior REBECCA publication [8], we referred to the characterization needs for Edge AI assurance. We now provide a full description and details of the means that can be used via a SMS.

The remainder of the paper is organized as follows. Section 2 presents the background. Section 3 discusses related work. Section 4 then describes the research methodology that we applied in the context of the conducted mapping study. Section 5 presents the results, whereas Section 6 discusses them. Section 7 deals with potential threats to validity of our work. Finally, Section 8 concludes the paper with a summary and outlook.

2. Background

We have divided the background of our work into two areas: **system assurance** and **Edge AI**.

System assurance refers to "*the set of activities to provide confidence that a system can be deemed dependable, e.g., safe and secure*" [20]. In general, it is based on the collection of evidence artifacts to show how a system's lifecycle *complies* with assurance standards and to substantiate why a system can be deemed *dependable* [21]. The former relates to compliance justification, e.g., regarding how the consistency of system requirements has been confirmed. The latter relates most often to technical risk reduction justification, e.g., to ensure that all the identified system hazards have been sufficiently and adequately addressed. This can be represented in a structured way in an *assurance case*. In both cases, there must exist *evidence* artifacts that support the declarations of compliance and of dependability. Figure 1 illustrates system assurance aspects and their relationships. It is also common to have to justify why a system manufacturer or component supplier has confidence in the suitability of the work conducted to ensure dependability, as it is impossible to fully demonstrate a specific dependability concern[2], e.g., system safety.

Assurance of systems is a complex activity [20]. Assurance standards are typically large textual documents that can consist of hundreds of pages and define thousands of compliance criteria. Ambiguity and inconsistency are common in these documents. System developers can easily face challenges when having to follow and apply assurance standards, manage large amounts of assurance evidence, and provide valid justifications of system dependability, among other difficulties. These difficulties can lead to assurance risks, which are conditions that can make a system developer incapable of (1) developing a system that complies with assurance standards and can be deemed dependable, (2) adequately collecting and managing assurance evidence and thus guaranteeing system dependability, or (3) making a third-party (e.g., an assessor) gain sufficient confidence in system dependability.

[2] *Concern* is defined as interest in a system relevant to one or more stakeholders [15].

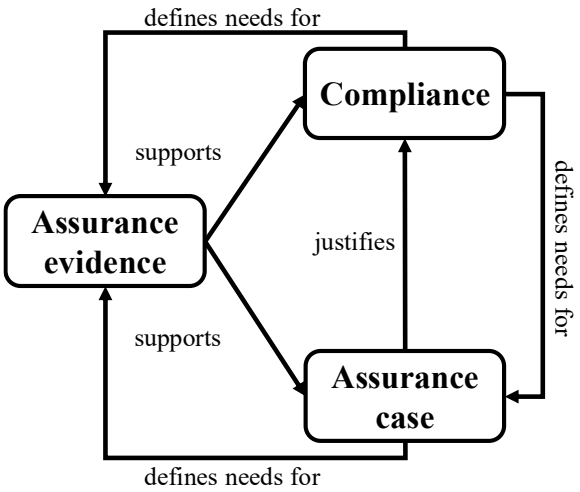


Figure 1. System assurance aspects and their relationships

Regarding **edge computing**, it is a distributed computing paradigm that brings computation and data storage closer to the edge of the network, closer to where data is generated and actions are taken [45], i.e., either directly on the device generating/collecting the data or on a nearby server (which is also referred as *fog computing* or *fog edge* [59]). Figure 2 illustrates the edge computing paradigm. By processing data locally, at or near the source, edge computing reduces latency and bandwidth consumption while enhancing real-time decision-making capabilities. This is particularly crucial for time-sensitive environments, such as autonomous vehicles, smart cities, and industrial automation.

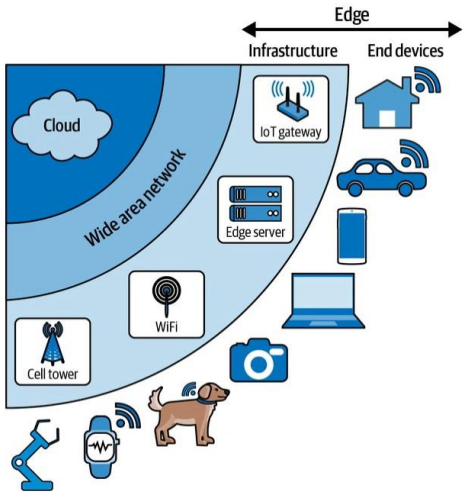


Figure 2. Edge computing paradigm (extracted from [69])

Edge computing is often interrelated to other terms such as *embedded systems*, *cyber-physical systems* (CPS) and *Internet of Things* (IoT). Embedded systems are specialized computer systems designed to perform specific tasks within larger systems or devices (e.g., wearable devices, industrial sensors, and healthcare monitors). They are the backbone of edge computing, enabling local data processing and analysis [58]. CPS are integrations of computation, networking, and physical processes. Embedded systems and networks monitor and control the physical processes, usually with feedback loops where physical processes affect computations and vice versa. Typically, IoT environments are composed of connected embedded systems that can collect and generate data (from cyber and/or physical sources), and can process it locally or not (i.e., IoT environments with or without edge capabilities) [17].

In turn, we refer to **Edge AI** as the *performance of AI computation on local devices close to the user (edge) and that are connected to some external server (e.g., a cloud)*, as defined in [76]. This means that AI algorithms are processed locally, either directly on the device or on a server near the device (Figure 3). The AI algorithms utilize the data generated by the devices themselves so devices can make independent decisions close to where the data is generated [69]. The cloud (or a private server) only participates when additional processing or storage is required [12,76].

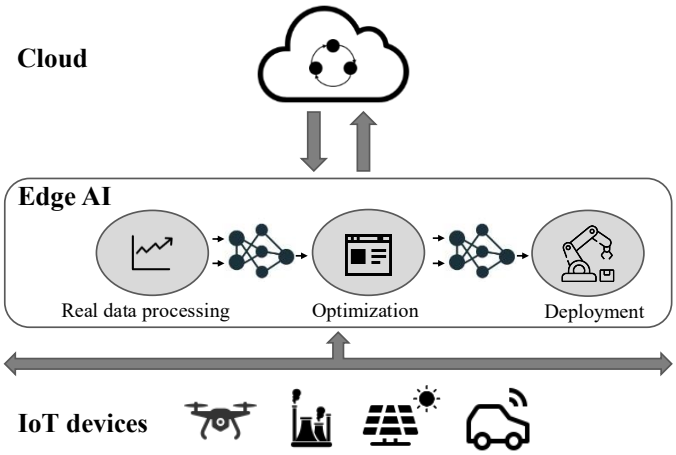


Figure 3. Edge AI (adopted from [67])

Examples of Edge AI applications include: healthcare monitoring, e.g., remote patient monitoring and personalized healthcare by analysing data from wearable devices and alerting healthcare providers about potential health issues; smart cities, e.g., data from sensors and cameras deployed throughout an urban area can power various smart city applications, including traffic management, and; autonomous vehicles, e.g., analysing sensor data in real-time for tasks such as object detection and collision avoidance. For the latter, we need to remark that we consider an autonomous system using AI an Edge AI system only if it is explicitly connected to a cloud (e.g., an autonomous car sharing data with a traffic external server). Thus, autonomous systems performing AI tasks in isolation (without the connection to a cloud) are not considered as Edge AI applications (e.g., two intelligent cars interconnected but not connected to an external server).

Finally, combining the above definitions, we can define **Edge AI assurance** as *the set of activities necessary to provide adequate confidence and evidence that an Edge AI system (local devices and external servers) satisfies given requirements*. Given this, in terms of *compliance management*, for Edge AI systems, it is needed to analyse how compliance progresses, possible gaps and interactions between safety and security [49]. Relevant standards and parts to comply with need to be identified. In addition, Edge AI-specific aspects to justify and justification structures to do it (e.g., according to usage context) need to be provided (i.e., *assurance case development*). The artefacts that contribute to developing confidence in the dependable operation of an Edge AI system and to showing the fulfilment of the requirements of one or several assurance standards (i.e., *assurance evidence*) need to be determined and evaluated [18].

Assuring Edge AI systems presents distinct challenges that differ fundamentally from those encountered in standalone edge computing or traditional AI assurance. This is due to the confluence of distributed infrastructure and intelligent functionality within constrained environments. Edge AI systems involve the execution of AI models (e.g., machine learning algorithms) on local devices that operate close to the data source and interact with external servers. Consequently, Edge AI assurance must address both the reliability of AI components and the dependability of edge environments, while

accounting for dynamic conditions such as model changes, runtime adaptation, and variability. Furthermore, characteristics inherent to edge settings—such as distribution, communication latency, real-time requirements, and resource constraints—lead to new limitations and methodological differences in assurance practices (e.g., safety cases or testing). Finally, unlike isolated AI or edge deployments, Edge AI systems necessitate multi-concern assurance strategies that integrate safety, security, privacy, and trustworthiness. For example, in edge environments, where devices are distributed, resource-constrained, and frequently connected to external networks, measures that strengthen security—such as encryption or access control—may inadvertently introduce latency or computational overhead that compromises safety-critical responsiveness.

Several initiatives exist to address edge and AI complexity and ensure their quality. The European Union's AI Act (*Regulation (EU) 2024/1689*) sets the requirements and obligations regarding specific uses of AI [24]. The International Telecommunication Union[3] has developed the AI standardization roadmap (*ITU-T Y.3000-series*) to assist in the development of AI related standards [44]. In turn, IEEE has defined the *IEEE P2805.3* standard for machine learning (ML) collaboration protocols on edge computing [36], the *IEEE P2802* standard for AI performance and safety evaluation for medical devices [34], and the *IEEE 1935* standard for edge/fog manageability and orchestration [37]. Finally, the International Organization for Standardization (ISO) has specified the ISO/IEC TR 24028:2020 report for AI trustworthiness [41] and the ISO/IEC TR 5469:2024 report for functional safety and AI systems [42]. However, despite these efforts, a unified and dedicated standard for Edge AI assurance remains elusive. In addition, there exist efforts aimed to assure individually edge and AI (e.g., AI assurance in general or hypervisor assurance [13,14,29,51]). However, they do not cover the Edge AI characteristics explained above in conjunction (e.g., smart devices connected to an external server).

3. Related work

To the best of our knowledge, no prior secondary study has dealt with Edge AI assurance. Nonetheless, prior publications have presented reviews and mapping studies whose scopes are related, but different, to our SMS. Several papers have analysed AI, edge, or assurance literature.

First, we are aware of systematic literature reviews (SLRs) on assurance in general. For example, Nair et al. [54] present an SLR about safety evidence. The results include a taxonomy (classification system) for different types of safety evidence as well as a review of the existing techniques for organizing and assessing them. Shukla et al. [66] present an SLR about system security assurance, emphasizing the risks in cyber-physical environments. This study systematically examines security requirements, metrics, assurance methods, and system environments to identify gaps and propose improvements (e.g., dynamic assurance approaches).

Second, some publications have reviewed dependability aspects in the edge (e.g., safety, security). Amiri et al. [4] present an SLR about the most frequent dependability concerns that play pivotal roles in resilience management of distributed environments (i.e., security, latency, and fault tolerance). Ashouri et al. [5] present a mapping study on quality attributes in edge computing for the IoT, e.g., time and resource utilization. This study also emphasizes the lack of empirical validation in actual IoT deployments. In turn, Bardis et al. [11] present approaches and techniques to improve IoT dependability in a non-systematic way. The authors analyse metrics for assessing software requirements' correctness to ensure software reliability. Another SMS is presented by Sanchez et al. [65], focusing on edge computing for CPS and emphasizing trustworthiness. It identifies challenges in integrating edge

[3] The United Nations specialized agency for information and communication technologies.

computing into CPS, particularly in ensuring privacy, availability, and real-time constraints. Also, Chinnasamy et al. [15] report a systematic review on data security and privacy requirements in edge computing, intending to evaluate protection and confidentiality standards. The review classifies different types of threats affecting edge devices and discusses how security technologies can mitigate them.

Third, a few reviews have been conducted about AI and safety-critical applications (e.g., avionics). Adler et al. [1] present a mapping study on the challenges of AI and its development that hinder its use in safety-critical applications. Among others, these include lack of explainability (AI operates as a black box) and real-time constraints (safety-critical systems require predictable response times, which AI may not always guarantee). In turn, Babeshko and Di Giandomenico [10] analyse systematically safety and cybersecurity assessment techniques for critical industries (e.g., hazard analysis and risk assessment). As a conclusion, the study advocates for developing a unified assessment technique that integrates both safety and cybersecurity considerations. Also, Mohseni et al. [52] present a safety-oriented categorization of ML techniques to provide guidance to improve dependability of ML design and development in safety-critical applications.

Fourth, there are reviews on AI assurance. For example, Neto et al. [57] present an SLR on safety assurance of AI-based systems. The study identifies five key strategies: black-box testing, safety envelopes, fail-safe AI design, white-box analysis combined with explainable AI, and lifecycle-based safety assurance. Stratigopoulos et al. [70] review the state of the art (e.g., fault injection techniques) for testing and reliability of hardware implementations of neuromorphic computing based on spiking neural networks. Tambon et al. [73] report on an SLR on how to certify ML based safety-critical systems considering robustness, uncertainty, explainability, verification, safe reinforcement learning and direct certification. In turn, Torens et al. [74] present a literature study on ML verification and safety for unmanned aircrafts (i.e., aircrafts with no human pilot on board). The study highlights that existing aviation safety standards do not fully accommodate ML-based automation, requiring new certification approaches. Moreover, Vyhmeister and Castane [75] provide a review on trustworthy AI technologies used in Industry 5.0. The authors identify gaps in industrial AI adoption, particularly in ensuring transparency, fairness, and accountability in industrial applications. Zhang and Li [82] systematically study the use of testing and verification of AI control software in safety-critical cyber-physical systems. The paper highlights difficulties in verifying AI behaviour, particularly in achieving repeatability and defending against common-cause failures. Finally, Zhang et al. [83] systematize how to evaluate the robustness of AI in safety-critical systems. This work categorizes eight evaluation approaches (e.g., adversarial robustness testing) and 16 metrics (e.g., adversarial perturbation sensitivity) aim to measure how well AI perform under various conditions.

In general, while several systematic reviews exist in the literature, the above publications primarily focus exclusively on edge computing, on AI, or on assurance. These reviews provide valuable insights into their respective domains. However, they do not address the combined interplay of Edge AI applications and assurance. Exploring these aspects in an interconnected manner provides a deeper and more comprehensive understanding of the challenges and opportunities in the field, offering a holistic perspective that these existing reviews do not capture.

Fifth, there exist also SLRs on computing areas that one could relate to Edge AI assurance, e.g., autonomous systems, intelligent systems, and self-adaptive systems. Nascimento et al. [56] present an SLR about the impact of AI on autonomous vehicle safety. The paper systematically maps AI safety research into six distinct categories, providing a structured understanding of how AI impacts autonomous vehicle safety (e.g., AI-based perception and sensing, decision-making and control,

human-AI interaction). In turn, a systematic review on security and safety of self-adaptive systems is presented by Pekaric et al. in [61]. It concludes that many existing approaches (e.g., hazard analysis) fail to consider the adaptation process when analysing security threats and safety hazards. It also highlights that safety and security are often treated separately, leading to gaps in comprehensive risk assessment. Similarly, Rajabli et al. [63] present an SLR on software verification and validation of safe autonomous cars. The study highlights difficulties in evaluating machine learning models used in autonomous cars decision-making. Also, Sun et al. [72] provide a survey on cyber-security of connected and autonomous vehicles with the aim of highlighting security problems and challenges, identifying also cyber-security standards for connected and autonomous vehicles. Finally, Gyllenhammar et al. [28] present an SLR on safety evidence for automated driving systems investigating design techniques, verification and validation methods, run-time risk assessment, and run-time (self-)adaptation. However, all these pieces of work do not fully align with our definition of Edge AI described in Section 2 (i.e., performing AI computation on local devices connected to some external server).

Sixth, we are aware of a non-systematic survey on security threats for edge computing platforms [80]. It details the most influential and basic attacks (e.g., attacks exploiting communication channels) as well as the corresponding defence mechanisms (e.g., restricting access) that have edge computing-specific characteristics. However, it does not consider AI executed in the edge.

4. Research method

A *Systematic Mapping Study* (SMS) is a means of identifying, categorizing and interpreting available research relevant to a particular topic area [48,62]. Individual studies contributing to a systematic mapping are called *primary studies*. A systematic mapping is a form of *secondary study*. The following subsections present the research questions, the data sources, the search string, the inclusion and exclusion criteria, the quality criteria, the study selection, and the data extraction and analysis.

4.1. Research questions formulation

The goal of this SMS is to identify the existing research on assurance for Edge AI by systematically selecting and reviewing published literature, structuring this field of interest in a broader, quantified manner. Note that a detailed understanding of the way assurance can be managed in the context of Edge AI requires an analysis of various aspects: assurance (e.g., compliance management, assurance case development, and assurance evidence management) and how AI is used at edge devices (e.g., dedicated hardware such as hypervisors). To further refine this objective, we formulated a set of Research Questions (RQs) to be answered by the selected primary studies:

- **RQ1:** What dependability concerns have been addressed? (e.g., safety, security, or privacy).
  This RQ allows us to assess existing efforts and identify areas requiring further attention to enhance the dependability of Edge AI.

- **RQ2**: In which domains has Edge AI assurance been applied? (e.g., healthcare or automotive)
  This RQ allows us to contextualize Edge AI relevance across industries.

- **RQ3:** What standards/regulations have been considered when assuring Edge AI? (e.g., DO-178C)
  This RQ allows us to investigate the alignment of existing research with regulatory and industry guidelines and/or norms. It addresses the relevance of Edge AI assurance practices to industry.

- **RQ4:** What types of assurance evidence have been managed? (e.g., requirements specifications or testing results)
  This RQ allows us to determine how dependability is substantiated to meet certain requirements, facilitating verification and certification processes.

- **RQ5:** What dependability justification techniques have been used? (e.g., safety cases)
  This RQ allows us to examine the formal techniques used to explain how Edge AI meets certain requirements.

- **RQ6:** What AI techniques have been considered? (e.g., ML algorithms)
  This RQ allows us to gather what AI techniques have been the most common for being assured at the edge.

- **RQ7:** What edge-specific characteristics have been considered? (e.g., the use of hypervisors or of accelerators, or network features)
  This RQ allows us to identify what characteristics of the edge environments have been considered for assurance.

4.2. Data source selection

Since assuring Edge AI covers interdisciplinary communities (assurance, AI, and edge computing), it is difficult to find an effective search string. For such purpose, based on our knowledge and prior experience [7, 19,20,54,71], we first performed an **exploratory search** over Google Scholar to identify proper keywords for the search string. These keywords were refined using trial and error. For example, we excluded terms whose inclusion did not yield additional studies (e.g., robustness). In addition, these searches provided us with an overview of existing literature and allowed us to identify candidate studies to be included as primary studies. The exploratory search was performed by all the authors.

After this exploratory search, we executed a hybrid search strategy [53], applying both a search in Scopus[4] and a snowball strategy [78]. This strategy is referred to as "**Scopus + BS*FS**" [53]: it first runs a search over Scopus, and then a (backward and forward) snowballing process is performed. Scopus is widely recognized as the leading scientific database due to its comprehensive coverage, rigorous indexing standards, and multidisciplinary reach. It contains peer-reviewed publications from top software engineering journals and conferences, including IEEE Xplore, ACM Digital library, ScienceDirect (Elsevier), and Springer research papers. In addition, Scopus has been widely used by prominent researchers in software engineering for systematic studies (e.g., [46][77][79]) and its coverage has been assessed as optimal when compared to these other databases [16].

For the snowballing strategy, we applied the guidelines indicated in [78] for backward and forward snowballing. We considered all the citations of each paper during backward snowballing. The source used to check forward references was Google Scholar because it offers the most extensive citation tracking capabilities. In addition, it captures a broader set of citing documents and more recent works (not yet indexed in Scopus). If any relevant study was identified, it was selected as a primary study and inserted in the snowballing process for another iteration. This hybrid search was primarily carried out by the first author. The second author reviewed and validated her outcomes. In cases of disagreement, the third author was consulted to arbitrate and facilitate consensus. Additionally, the third author conducted an independent random inspection to further validate the performance of the search strategy.

[4] www.scopus.com

Finally, Google Scholar Alerts (e.g., “AI AND assurance AND edge”) were continuously analysed to become aware of any publication on the topic emerging during the writing process, i.e., after the search was performed. Forward references were also checked.

4.3. Search string

The final search string was:

[part I]

({edge artificial intelligence} OR {edge ai} OR "intelligent edge" OR {edge intelligence}
OR (
({edge computing} OR "edge device" OR "edge system" OR "edge solution" OR {at the edge} OR {on the edge} OR {in the edge} OR "cloud edge" OR "edge cloud" OR "fog edge" OR "edge fog"
OR (({cloud} OR {fog}) AND {edge} AND ({iot} OR "internet of things")
)
)
AND ({ai} OR {intelligence} OR learn OR {ml} OR "autonomous s*" OR "autonomous device" OR "neural network" OR "smart s*" OR "smart device" OR "intelligent s*" OR "intelligent device" OR robot*)*
)
)

[part II]

AND (
(assur OR {compliance} OR {comply} OR {complies} OR {compliant} OR qualif* OR certif* OR dependab* OR trustworth**

)

The first part of the search string captures keywords related to Edge AI. We considered several keywords in addition to “edge artificial intelligence”. These additional keywords capture terms that are sometimes related to ‘edge’ as described in Section 2 (e.g., IoT, fog) and concepts that could share some properties with Edge AI (e.g., smart device). The second part concerns system assurance. We also included keywords representing activities that share the underlying principles with system assurance (e.g., certification and qualification). We did not include general terms (e.g., ‘safety’ or ‘security’) since they retrieved an unmanageable number of studies.

To retrieve as many potential studies as possible, we accommodated the search string using the syntax rules provided by the Scopus search guide[5]. We used the asterisk * to include the different variants of the same keyword (e.g., certified vs certification, dependable vs dependability). In addition, we used brackets () to use Boolean operators (i.e., AND, OR). Finally, we enclosed the terms in braces {} to search an exact phrase, including any stop words, spaces, and punctuation (e.g., {edge artificial intelligence}), and in double quotation marks “” to search an approximate phrase, i.e., punctuation is ignored, and plurals are included (e.g., “edge device”).

4.4. Inclusion and exclusion criteria

We defined the following inclusion and exclusion criteria to select relevant studies. The basic inclusion criterion was to include studies whose work is focused on Edge AI assurance. We would like to highlight that we were flexible on how Edge AI was defined in the study (e.g., use of different terminology/synonyms). However, we were strict in terms of assurance meaning that we only included studies where assurance is clearly explained or defined. For such purpose, we used RQ3 (i.e., standards),

[5] https://schema.elsevier.com/dtds/document/bkapi/search/SCOPUSSearchTips.htm

RQ4 (i.e., assurance evidence) and RQ5 (i.e., dependability justification techniques) as inclusion criteria, selecting only those studies that provided answers to them.

We also applied the following exclusion criteria, filtering out publications that matched any of the criteria:

1. The study is not related to Edge AI assurance, or it merely mentions Edge AI assurance terms in a general manner. In particular, studies were excluded if their relevance to Edge AI assurance could not be established through explicit descriptions.
2. The study is not presented entirely in English.
3. The study presents some type of review (e.g., survey, SMS) but does not deal with outcomes of a particular research work.
4. The study is a book cover[6], a tutorial, a project deliverable or a poster publication.
5. The study is not electronically available (even after contacting the authors).
6. In case several studies refer to the same research work, all studies except the latest and most complete version are excluded.

We did not apply any restriction with respect to the publication date. However, since Edge AI assurance is a novel topic, studies were expected to be recently published (e.g., since 2018).

4.5. Quality criteria

In addition to the inclusion and exclusion criteria, each selected study was assessed based on a set of quality assessment questions. Although a quality assessment is not mandatory in SMSs [62], this is important for checking if studies could potentially answer the research questions in a clear and scientific manner. In addition, this is important for interpreting and synthesizing the data extracted from the selected studies [48]:

1. Does the study include sufficient data to infer how Edge AI assurance can be performed?
   This question evaluates whether the study presents adequate data and methodological insights to enable a meaningful assessment of the presented Edge AI, e.g., it provides information about evidence types, edge-specific characteristics or standards considered.
2. Does the study provide any implementation, formalization or validation to support the assurance?
   This question evaluates whether the study presents concrete empirical information that substantiate the presented assurance. It is important to note that we used the term "*validation*" in a broad sense (i.e., not necessarily implying validation in a controlled environment such as a controlled experiment).

These questions allowed us to manifest the extent to which Edge AI assurance is described and to ensure a certain level of maturity for the studies included in the SMS.

[6] Book chapters were not excluded.

4.6. Study selection

As described in Section 4.2, the SMS was conducted using different types of searches (see Figure 4[7]). **First**, we performed an *exploratory search* in Google Scholar. This search was done to identify proper keywords and candidate studies to be included as primary studies. Concretely, we identified 13 primary studies.

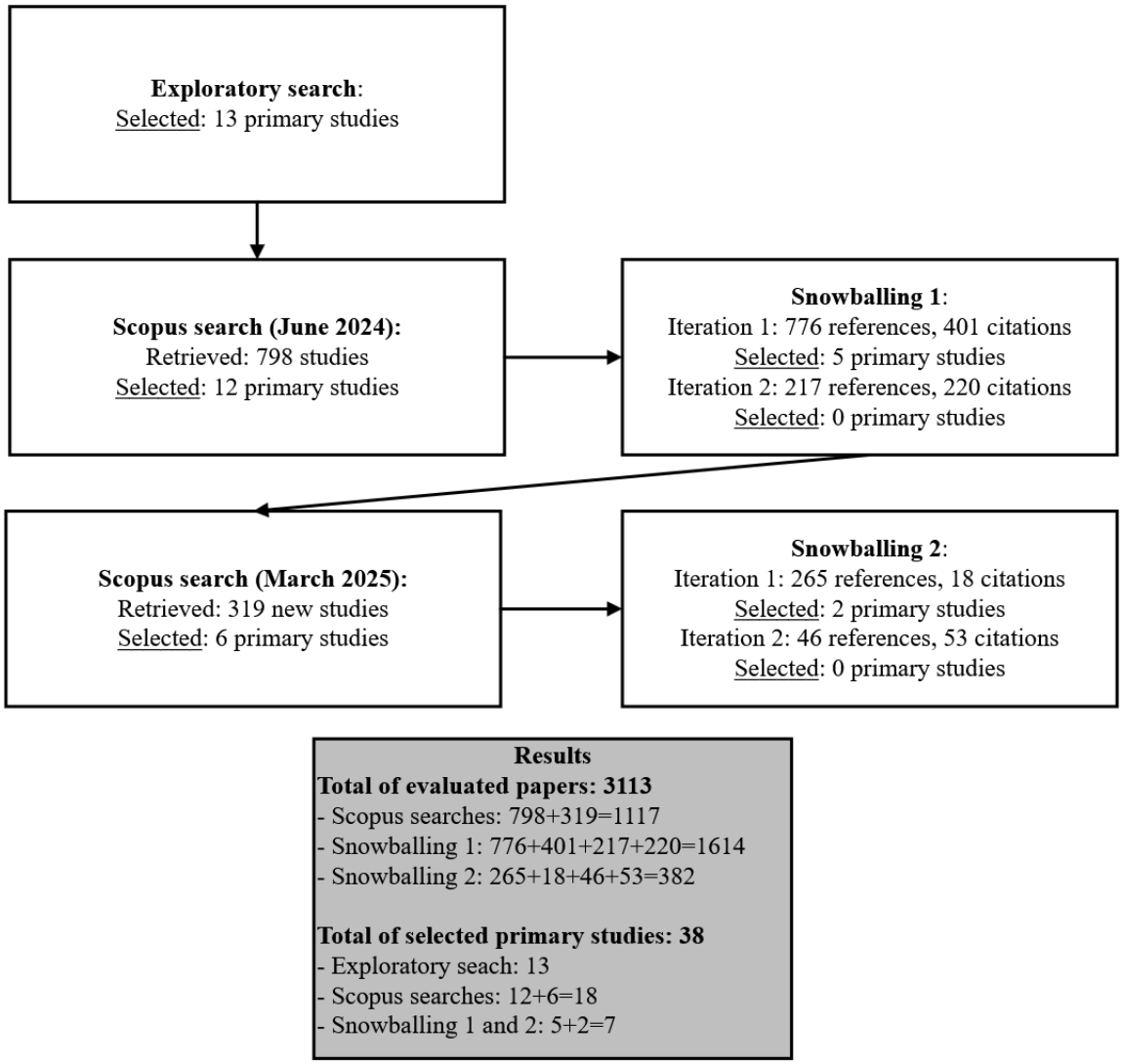


Figure 4. Study selection process

Second, in June 2024, we performed a *search in Scopus* using the defined search string (Section 4.3). This resulted in a total of 798 studies. Then, each study was reviewed to determine its relevance for the mapping. This was accomplished based on the defined exclusion criteria. First, the title and abstract of each retrieved study were analysed to check whether the studies dealt with Edge AI assurance (i.e., application of Exclusion Criteria 1–5). When the information from the title and abstract was not sufficient to decide whether to include the study, the full text was reviewed. Second, duplicated studies were removed (i.e., Exclusion Criterion 6). At the end, we obtained a total of 12 additional new primary

[7] The arrows indicate the chronological order in which the searches were conducted.

studies (see Figure 4). Altogether, combining both searches, we obtained a set of 25 primary studies (i.e., 13 studies from the exploratory search and 12 from Scopus).

Third, we performed the backward and forward snowballing process over these 25 primary studies applying the guidelines indicated in [77]. For backward snowballing, we analysed the literature cited in the background or related work section of each of these 25 primary studies (i.e., 776 references). For forward references, we used Google Scholar since it is the tool that performs better for this purpose [77]. We obtained a total of 401 citations[8]. Then, we applied the same filtering process described for the Scopus search over the obtained references and citations (i.e., read title and abstract, apply exclusion criteria, read the full paper if needed, and remove duplicated studies). As a result, we obtained five new primary studies. We then repeated the snowballing process over these new five studies and analysed 217 backward references and 220 citations. No new primary studies were found, and the snowballing process was finished (i.e., second iteration).

Fourth, since the process of analysing and reporting the results took longer in time than expected, in March 2025, we decided to conduct *a new search in Scopus* to ensure that our SMS results were as up to date as possible. In this second Scopus search, we only searched for studies published between June 2024 to March 2025, retrieving a total of 319 new studies. These were filtered as described above and we selected six new primary studies. **Fifth**, these studies were introduced in a *new snowballing process*. In the first iteration, we analysed 265 references and 18 citations from which we selected two new primary studies. Again, we repeated the snowballing process over these two new studies, analysing 46 references and 53 citations. No new primary studies were identified and thus this second iteration was finished.

The whole selection process outlined above resulted in 38 primary studies to be included in the mapping (1.22% from the 3113 total of studies evaluated): 13 from the exploratory search, 18 from the searches in Scopus and 7 from the snowballing processes (see Figure 4). Appendix A includes the complete information of the selected primary studies (i.e., tittle, authors, publication date and venue) sorted by order of appearance during the search.

The study selection process was carried out mainly by the first author and checked by her co-authors. More precisely, the co-authors (1) randomly reviewed selected studies to ensure consistency of the process, (2) performed exploratory searches to ensure that relevant studies were included, and (3) reviewed the correct application of the inclusion and exclusion criteria. All disagreements were resolved through discussion.

4.7. Data extraction and data analysis

A data extraction process was applied to each of the 38 primary studies, with the goal of answering the research questions defined in Section 3.1. For this purpose, we designed a data extraction template (a spreadsheet) to capture and store the relevant information. Appendix B includes an excerpt. All the information about the data extracted from all the studies can be found in [9]. In detail, we extracted:

1. Bibliographic information, i.e., title, authors, type of venue (e.g., conference, journal), and complete reference of the study.

2. The dependability concerns that have been addressed when assuring Edge AI (RQ1).

[8] We did not establish a limit over the returned citations.

3. The application domain in which the system under assessment was used (RQ2).
4. The standards that have been considered for Edge AI assurance (RQ3).
5. The assurance evidence that has been managed (RQ4).
6. The dependability justification techniques that have been used (RQ5).
7. The AI techniques that have been considered for assurance (RQ6).
8. The edge-specific characteristics that have been considered (RQ7).

For research questions RQ1, RQ3, RQ4, RQ5, RQ6 and RQ7, data was extracted by first creating an initial list of categories (e.g., safety and security for RQ1) based on our knowledge and experience about the topic [19,20]. Once data extraction started, each study was then thoroughly analysed and extracted data was assigned to a category based on content analysis techniques [31]; if new categories were identified, they were added to the list. Finally, for RQ2, we included each identified domain in which the system under assessment was used by analysing the content of each study. Again, throughout this analysis, similar domains might be merged. This may imply the reassignment of already analysed studies.

After completing the extraction, we proceeded to analyse and synthesise the data. In general, we performed both quantitative and qualitative analyses to classify the extracted data. We used descriptive statistics to analyse the results (i.e., frequency counts) to those research questions where lists were created (RQ1, RQ3, RQ4, RQ5, RQ6, and RQ7). To simplify the synthesis of the extracted data, we used descriptive techniques to summarize them, e.g., graphics and tabular descriptions. For RQ2, we relied on extensive qualitative analysis by examining the full text of each study. If extraction and/or analysis inconsistencies and mistakes were discovered, they were resolved through discussion.

5. Results

This section presents the results from the mapping, answering each research question individually based on the extracted data. Figure 5 shows the temporal distribution of the 38 primary studies by publication year (i.e., from 2019 to 2025). As can be seen, the topic of Edge AI assurance is recent and the number of published studies is increasing over time, with a peak in 2023.

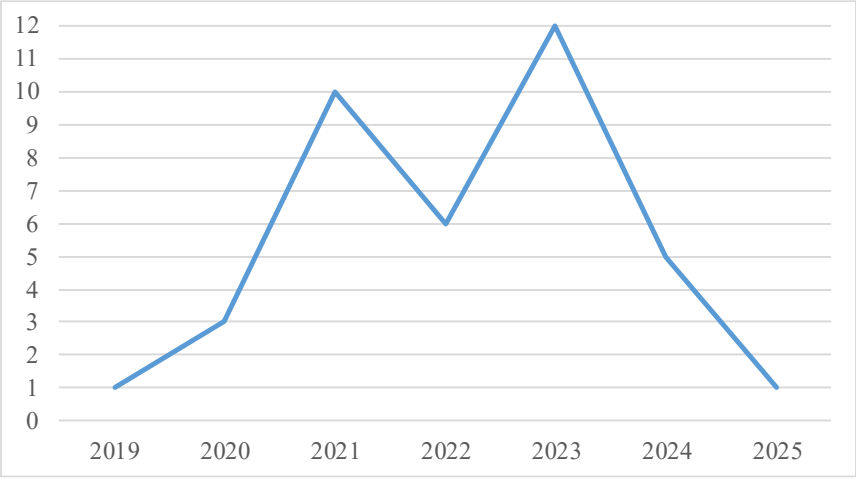


Figure 5. Distribution of primary studies by publication year

For each RQ, we present complementary visual and textual results derived from the findings. While figures offer a global and aggregated overview of distributions, text and tables (if needed) provide the detailed results for each primary study in each RQ.

5.1. RQ1: What dependability concerns have been addressed?

This section presents the dependability concerns that have been addressed for assuring Edge AI (RQ1). We identified several concerns addressed by the selected primary studies (Figure 6):

- *Safety:* absence of catastrophic consequences on the user(s) and the environment [6].
  Studies ID2, ID7, ID8, ID10, ID11, ID14, ID15, ID20, ID21, ID26, and ID29.

- *Security:* being protected from unauthorized access, remain accurate and unaltered, and being accessible when needed [39].
  Studies: ID6, ID9, ID13, ID16, ID22, ID24, ID28, ID31, ID33, and ID38.

- *Trustworthiness*: assurance that a system will perform as expected [6].
  Studies: ID1, ID19, ID30, ID31, ID36, and ID37.

- *Dependability* (in general): the ability to avoid service failures that are more frequent and more severe than is acceptable [6].
  Studies: ID3, ID4, ID18, ID25, and ID34.

- *Trust:* accepted dependence [6].
  Studies: ID23 and ID35.

- *Accuracy:* closeness of computations or estimates to the exact or true values that the statistics were intended to measure [27].
  Study: ID5.

- *Effectiveness:* the extent to which planned activities are realized and planned results achieved [27].
  Study: ID17.

- *Performance:* the efficiency, effectiveness, and accuracy with which a system, process, or individual completes a task [43].
  Study: ID12

- *Privacy*: protection of sensitive data [25].
  Study: ID32.

It is important to note that only one study (ID31) addresses two different concerns in combination (security and trustworthiness). Finally, one study (ID27) was identified addressing non-functional properties in general. It names some examples of these properties (e.g., confidentiality, fairness, and explainability) but it does not provide details on how they are addressed individually.

Note that we have adopted a broad interpretation of dependability concerns including concerns that are not typically classified as dependability attributes in established standards (e.g., accuracy, effectiveness, and performance). We decided to do this to properly reflect the state of literature since all the identified concerns are, in some way, connected to dependability, as reflected in foundational dependability references such as [6].

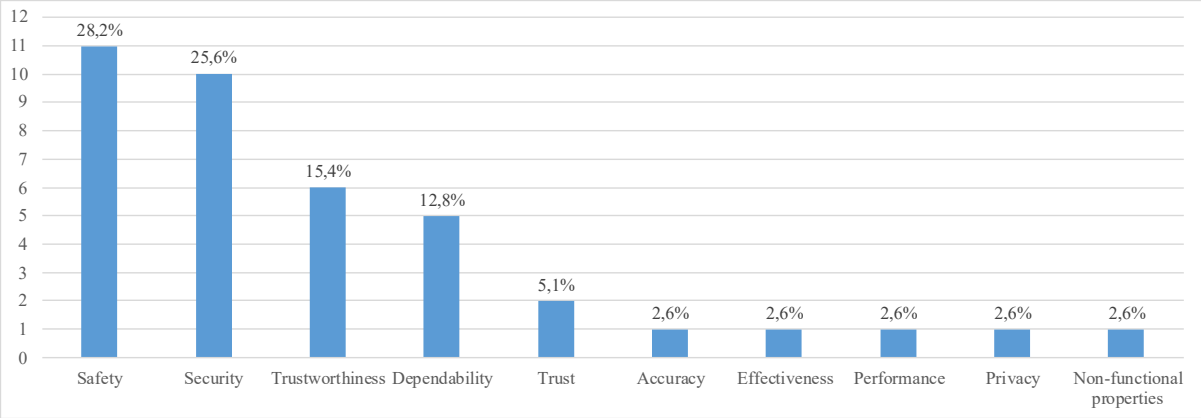


Figure 6. Distribution of primary studies according to the dependability concern addressed

5.2. RQ2: In which domains has Edge AI assurance been applied?

This section gives insights into the application domains identified in the studies (i.e., RQ2). These domains are (Figure 7):

- *Industry 4.0*: intelligent digitalised industrial environments (also named *Industrial Internet of Things* [68]).
  Studies: ID1, ID11, ID15, ID21, ID24, ID25, and ID30.

- *Automotive* (cloud-connected autonomous vehicles): intelligent self-driving cars.
  Studies: ID8, ID14, ID17, and ID29.

- *Smart city*: digitalised urban area.
  Studies: ID3 and ID22.

- *Unmanned Aerial Vehicles* (UAV): intelligent aircrafts with no human pilot, crew, or passengers on board.
  Study: ID7.

- *Space*: intelligent devices sent up into space.
  Study: ID2.

- *Healthcare*: intelligent devices that collect and process data to enhance diagnosis, treatment, and patient care.
  Study: ID32.

Note that the above list of domains includes only those cited in the studies and how they are explicitly named in them. In addition, we found 11 studies (28.9%) that refer to technologies (i.e., CPS, IoT, video capturing/streaming, and microservices) that could be applied to any domain (ID9, ID10, ID12, ID16, ID18, ID19, ID20, ID26, ID28, ID35, and ID37). Finally, we found 11 studies (28.9%) where Edge AI assurance is discussed in a general context, without association to any specific application domain (ID4, ID5, ID6, ID13, ID23, ID27, ID31, ID33, ID34, ID36, and ID38).

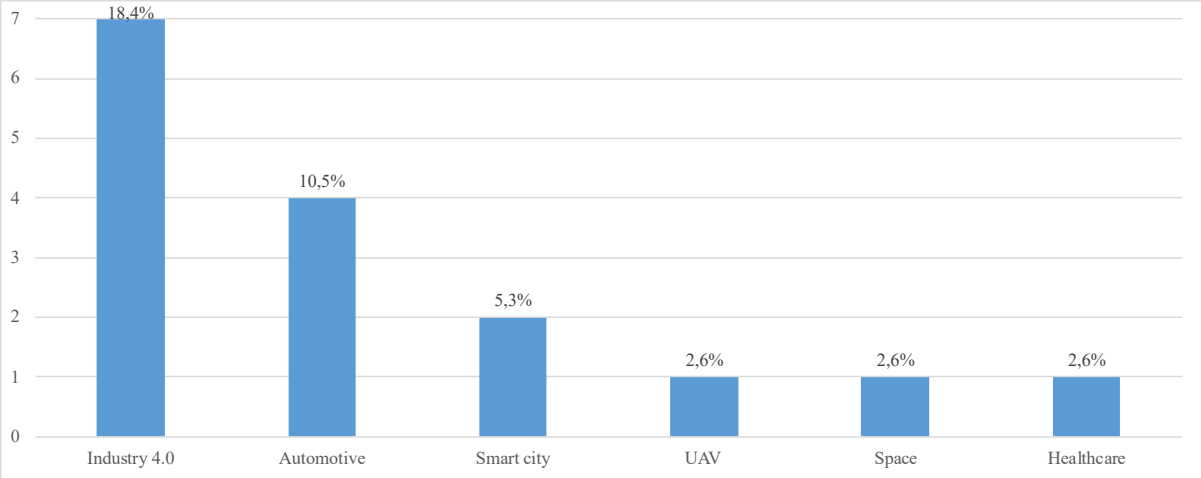


Figure 7. Distribution of primary studies by domain

5.3. RQ3: What standards have been considered when assuring Edge AI?

This section summarizes the standards/regulations that have been considered when assuring Edge AI (RQ3). It is worth noting that we identified (and thus included) all the standards or regulations considered in each study, not only the ones dedicated to dependability assurance (e.g., safety or security, such as ISO26262). Concretely, we identified eight different standards and regulations (Figure 8):

- *IEEE 802.1 TSN*: standard for deterministic connectivity through IEEE 802 networks [33].
  Studies: ID1and ID15.

- *ISO 26262:* standard for the functional safety of road vehicles [40].
  Studies: ID8 and ID18.

- *DO-333:* a supplement of the standard DO-178C that provides guidance on the use of formal methods to produce verification evidence for certifying safety-critical airborne software [64].
  Study: ID7.

- *oneM2M:* standard for IoT, providing the foundations for scalable and interoperable systems [60].
  Study: ID12.

- *ETSI C-ITS:* standard focused on developing intelligent transport systems that enable direct communication between vehicles and infrastructure to enhance road safety, traffic efficiency, and environmental sustainability [23].
  Study: ID17.

- *ANS/ITSDF B56.5:* standard establishing safety requirements, standardizing principal dimensions for interchangeability, and setting test methods for powered and non-powered industrial trucks [3].
  Study: ID21.

- *IEEE 802.15:* standard for wireless specialty networks [35].
  Study: ID28.

- *Legal and ethical regulations:* established frameworks to, for example, ensure data privacy, e.g., GDPR (*General Data Protection Regulation*) [25].
  Study: ID32.

Finally, 28 studies (73.7%) do not consider any standard (ID2, ID3, ID4, ID5, ID6, ID9, ID10, ID11, ID13, ID14, ID16, ID19, ID20, ID22, ID23, ID24, ID25, ID26, ID27, ID29, ID30, ID31, ID33, ID34, ID35, ID36, ID37, and ID38).

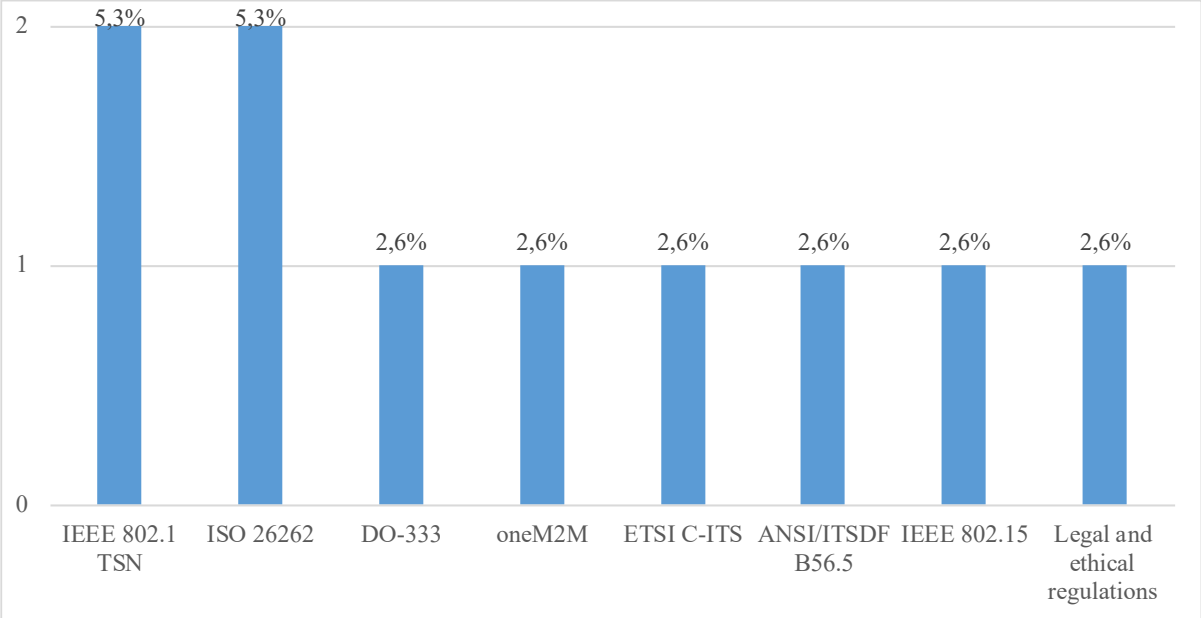


Figure 8. Distribution of primary studies according to the considered standards

5.4. RQ4: What types of assurance evidence have been managed?

This section provides insights into the types of assurance evidence to substantiate conformance to certain requirements (e.g., safety or security requirements) (RQ4). As reference framework, we used the evidence taxonomy presented in [54] to identify what evidence is considered by the selected studies. Since different studies had information at different abstraction levels (i.e., studies providing more details than others), we denote the lowest abstraction level identified for each evidence type. Figure 9 depicts the distribution of primary studies according to the identified evidence types.

As described in Section 4.7, the identification of evidence types was performed through content analysis during full-text review, guided by our domain knowledge and the taxonomy[9] established in [54]. That work also complements the taxonomy with examples of techniques (e.g., UML diagrams for design specification), artefacts (e.g., hazard logs for hazard specification), or information (e.g., failure frequency for reused component specification) that can be considered as or used to for the identification of evidence types (see Appendix B of [54]). Thus, we used this as a method to evidence identification and assignment. However, since the reference taxonomy was developed only for safety in critical systems in general [54], we have adapted and expanded it to reflect the current needs of Edge AI assurance. We included 'runtime verification results' as a new type of evidence. For understanding, Appendix C provides the taxonomy and the definition of each type of evidence found. Additionally, each evidence type is exemplified with content from the primary studies.

[9] The third author of this paper is the second author of [54].

In addition, Table 1 presents the evidence types identified in each primary study. We would like to highlight that most of the studies provide more than one type of evidence.

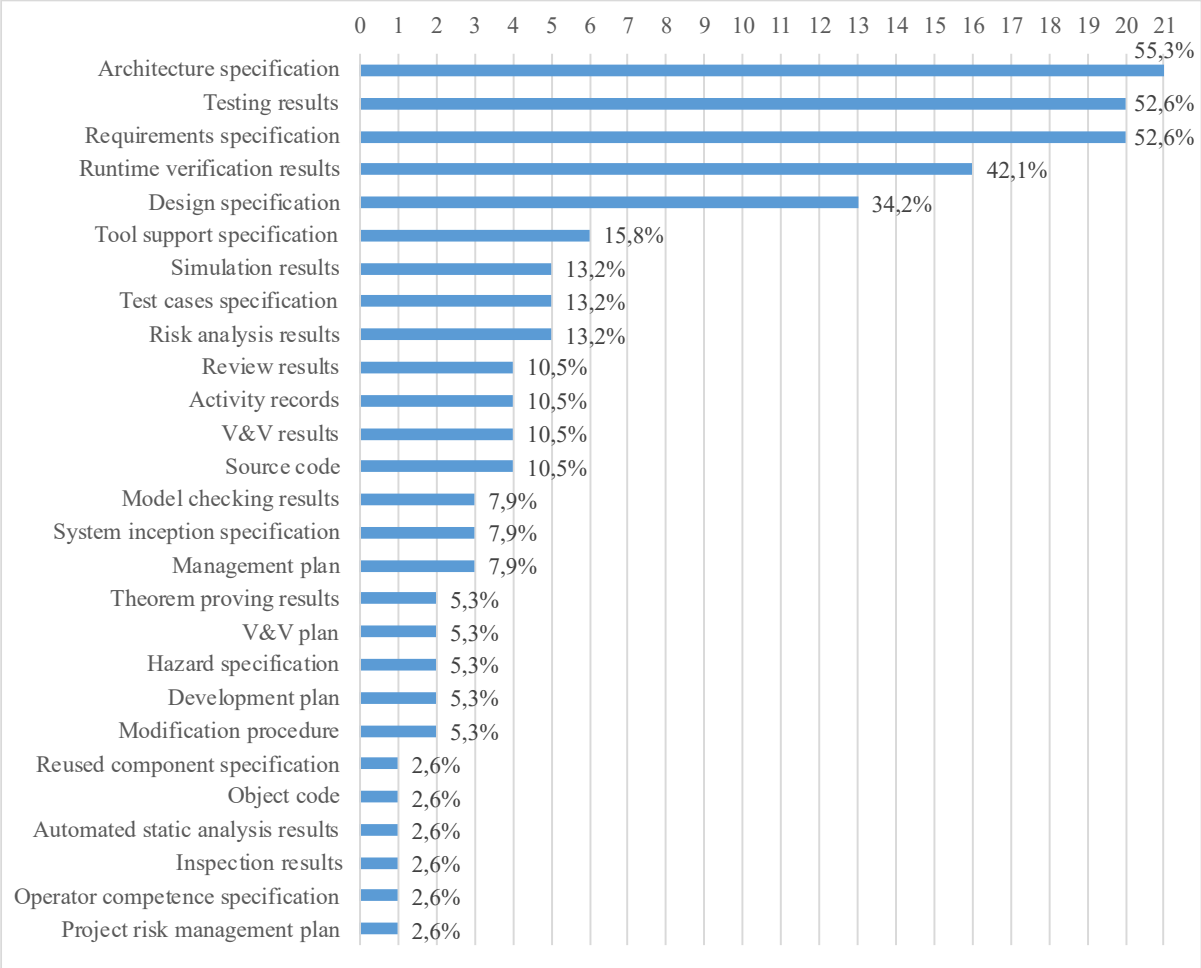


Figure 9. Distribution of primary studies according to the evidence types

Table 1. Studies considering each type of evidence

Type of evidence	Studies
Architecture specification	ID1, ID3, ID4, ID12, ID13, ID15, ID16, ID17, ID18, ID20, ID23, ID25, ID27, ID28, ID29, ID30, ID31, ID32, ID33, ID36, ID37
Testing results	ID1, ID2, ID4, ID8, ID10, ID13, ID14, ID15, ID16, ID18, ID20, ID21, ID25, ID26, ID29, ID31, ID34, ID35, ID37, ID38
Requirements specification	ID2, ID7, ID9, ID10, ID11, ID13, ID14, ID15, ID16, ID20, ID21, ID26, ID29, ID31, ID32, ID33, ID35, ID36, ID37, ID38
Runtime verification results	ID1, ID4, ID5, ID10, ID13, ID14, ID16, ID20, ID21, ID22, ID23, ID24, ID25, ID26, ID30, ID38
Design specification	ID13, ID16, ID22, ID24, ID25, ID29, ID30, ID31, ID32, ID35, ID36, ID37, ID38
Tool support specification	ID8, ID13, ID16, ID19, ID21, ID34
Simulation results	ID6, ID21, ID22, ID28, ID30
Test cases specification	ID2, ID10, ID13, ID14, ID35
Risk analysis results	ID11, ID18, ID20, ID22, ID23
Review results	ID2, ID7, ID11, ID13
Activity records	ID2, ID9, ID13, ID16
V&V results	ID9, ID15, ID33, ID34

Source code	ID9, ID32, ID35, ID36
Model checking results	ID7, ID10, ID21
System inception specification	ID19, ID20, ID29
Management plan	ID18, ID32, ID37
Theorem proving results	ID7, ID10
V&V plan	ID9, ID33
Hazard specification	ID11, ID21
Development plan	ID29, ID36
Modification procedure	ID29, ID37
Reused component specification	ID9
Object code	ID9
Automated static analysis results	ID13
Inspection results	ID13
Operator competence specification	ID26
Project risk management plan	ID32

5.5. RQ5: What dependability justification techniques have been used?

In this section, we present the techniques that are used to justify dependability in the selected studies (RQ5). As in RQ4, we use the techniques presented in [54] as reference of existing justification techniques[10] (Figure 10).

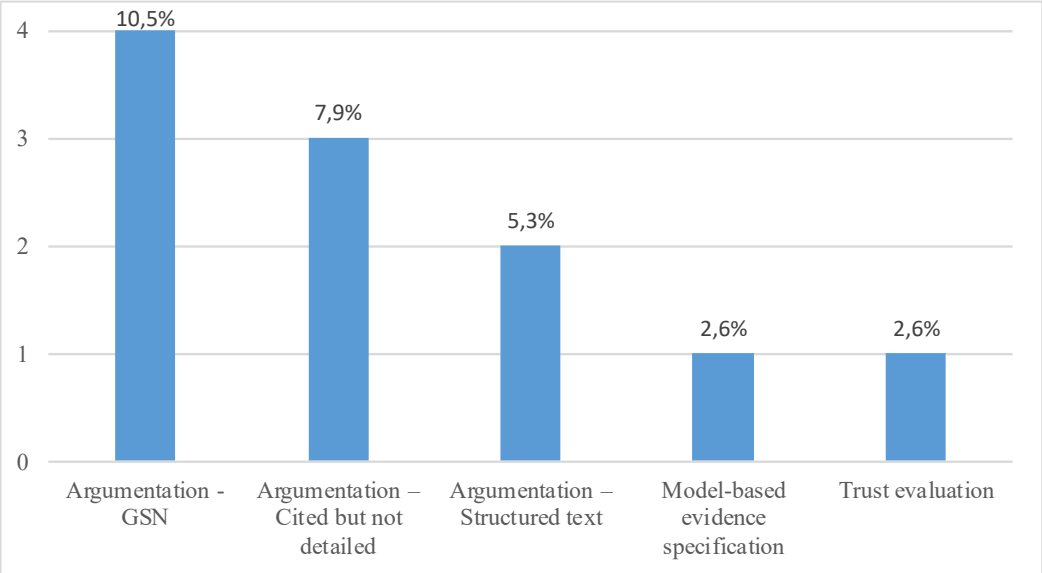


Figure 10. Distribution of primary studies according to the dependability justification techniques

From [54], we found two different techniques: argumentation and model-based evidence specification. On the one hand, argumentation is a technique that explains the reasons why a system is considered to be acceptably dependable (e.g., acceptably safe). The argumentation can be expressed either *graphically* (e.g., using *Goal Structured Notation*, GSN) or *textually* (e.g., structured text). In

[10] In [54], the dependability justification techniques are named '*techniques for structuring safety evidence*' with the same meaning.

GSN, the claims of the argument are documented as goals and items of evidence are documented in solutions. Examples include the assurance argument in ID2 for wildfire alert component, the GSN-based interface that facilitates the selection of relevant solutions in ID7, the assurance case contract for IoT systems in ID11, and the safety cases for smart factories in ID21. For the structured text, argumentation is presented using certain structure to regular prose (e.g., indentation, numbering, different fonts) to explicitly denote the parts of the argument (e.g., how to train the AI classifier in ID10 and the assurance evaluation in ID16). However, we also found studies that do not provide full details about the type of argumentation used despite argumentation is explicitly mentioned, e.g., "*to make a convincing argument that it satisfies all the rules expected*" (extracted from ID26), "*safety arguments are iteratively built and traced to the SRs/RSRs on a traceability matrix*" (extracted from ID29) and "*At the level of each functionality, constraints on arguments and outputs can be specified*" (extracted from ID33). On the other hand, model-based evidence specification refers to characterizing the structure of evidence using models of any kind (e.g., UML meta-models, entity-relationship diagrams, or process models), e.g., the process model about risk management in ID32.

Like in RQ4, the techniques provided in [54] do not cover all the needs for assuring AI at the edge (e.g., trustworthiness of AI algorithms). This is the case of ID28, where authors provide "*an evaluation of trust"* to justify the dependability of the Edge AI system. In particular, the proposed mechanism combines the concept of distributed (edge nodes) and centralized (central authority) trust management along with time-driven and event-driven trust computations.

Finally, we would like to highlight that 27 studies (71.1%) do not provide any justification technique (IDI, ID3, ID4, ID5, ID6, ID8, ID9, ID12, ID13, ID14, ID15, ID17, ID18, ID19, ID20, ID22, ID23, ID24, ID25, ID27, ID30, ID31, ID34, ID35, ID36, ID37, ID38).

5.6. RQ6: What AI techniques have been considered?

This section outlines the AI techniques considered in ensuring Edge AI (RQ6). We identified five main techniques (Figure 11):

- *Machine learning*: a type of AI that enables computers to learn patterns from data and make decisions without explicit programming [69].
  Studies: ID1, ID2, ID10, ID13, ID15, ID26, ID27, ID28, ID30, ID32, ID35, ID36, and ID37.

- *Deep learning*: a subset of machine learning that uses artificial neural networks to process and learn from large amounts of data [69].
  Studies: ID3, ID5, ID12, ID14, and ID25.

- *Federated learning*: a decentralized approach to machine learning where multiple devices collaboratively train a model without sharing raw data [69].
  Studies: ID19, ID24, ID31, and ID38.

- *Pathfinder*: AI-based search algorithms to find an optimal path from a specific start waypoint to a goal waypoint (e.g., A* search algorithm [38]).
  Study: ID7.

- *Convolutional neural networks* (CNNs): deep learning models designed for automated feature extraction and pattern recognition in structured data, such as images.
  Study: ID34.

We want to highlight that we report the AI techniques in the way they are considered in the studies. We acknowledge that deep learning and federated learning are subtypes of machine learning, as well as CNNs are types of deep learning models. However, we treat them separately to present the actual content of the studies.

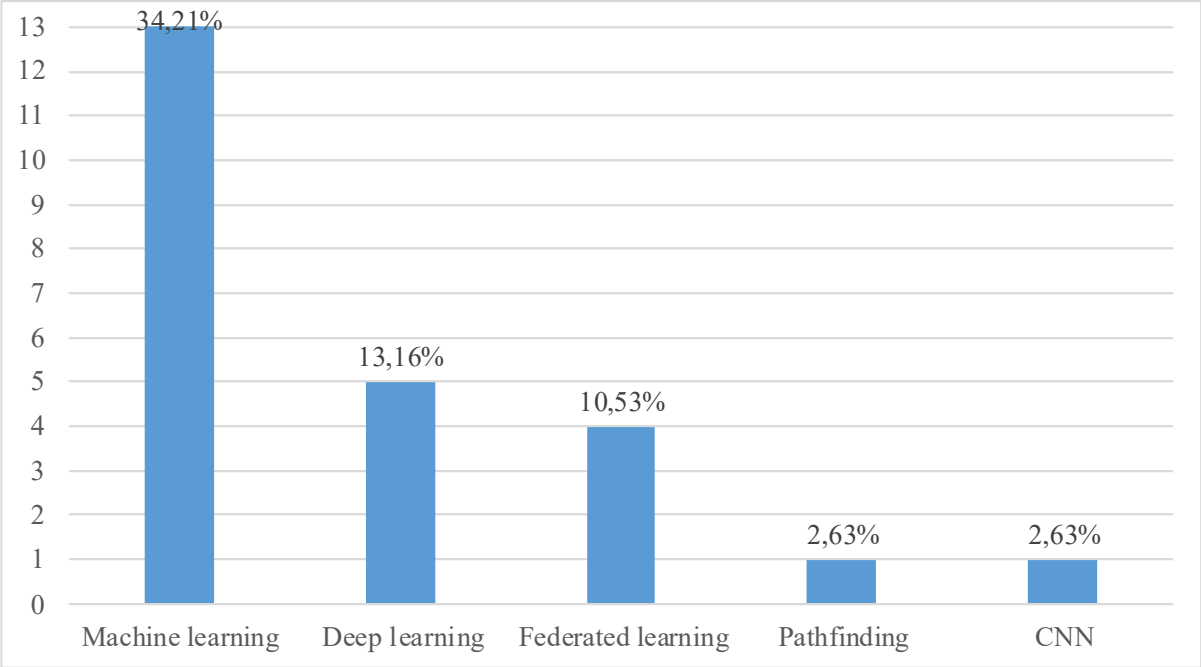


Figure 11. Distribution of primary studies according to the considered AI techniques

Finally, we also find that 14 studies (37%) do not specify any particular AI technique (ID4, ID6, ID8, ID9, ID11, ID16, ID17, ID18, ID20, ID21, ID22, ID23, ID29, and ID33). Instead, they refer to AI models or algorithms in general running in edge devices. Basically, most of these studies do not specify an AI technique since they focus on assuring other aspects of Edge AI, such as the hardware devices or the network, assuming that AI is involved.

5.7. RQ7: What edge-specific characteristics have been considered?

This section reports the results of RQ7 about the edge-specific characteristics that have been considered for assurance (Figure 12). We classify the results in the following categories:

- *End node*: assurance is provided for either the software (SW, e.g., AI algorithms) or the hardware (HW, e.g., AI accelerators) of the end nodes of the Edge AI infrastructure (e.g., smart device).
- *Network*: assurance is provided for the communication channel between the elements of the Edge AI infrastructure (e.g., communication between the cloud and the end nodes).
- *Fog infrastructure*: assurance is provided for the intermediate layer between the end nodes and the cloud (e.g., intermediate server).
- *Edge AI infrastructure*: assurance is provided for the complete Edge AI infrastructure, i.e., end nodes, network, cloud, and intermediate servers (if existing).

Table 2 details the primary studies for each edge-specific characteristic.

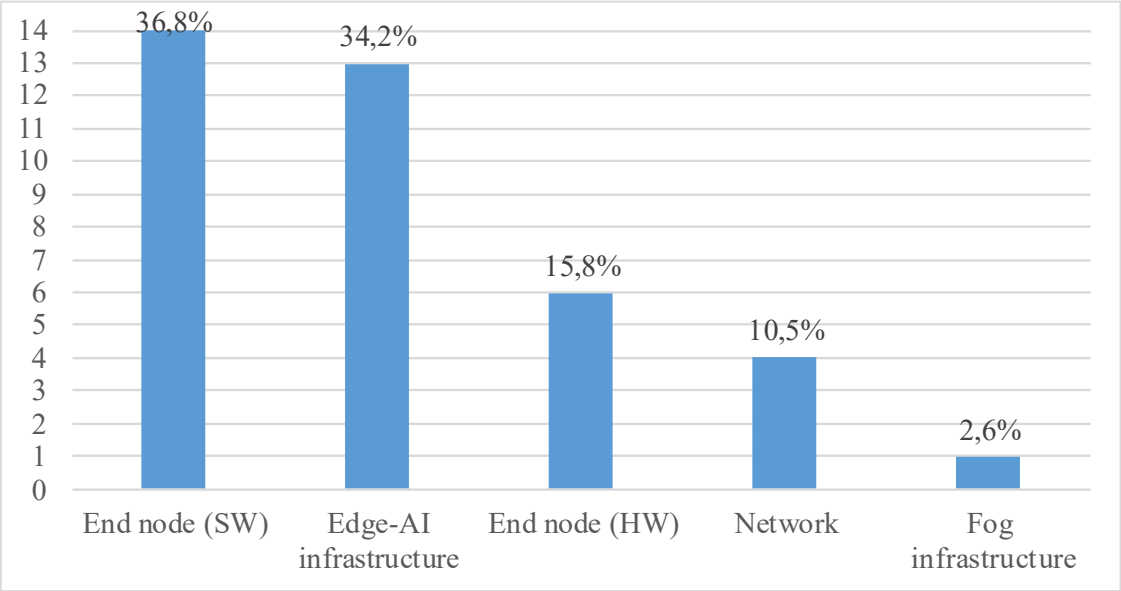


Figure 12. Distribution of primary studies according to the edge-specific characteristics considered

Table 2. Studies considering each edge-specific characteristic

Edge-specific characteristic	Studies
End node (SW)	ID2, ID3, ID5, ID6, ID7, ID10, ID11, ID12, ID13, ID14, ID28, ID31, ID32, ID38
Edge AI infrastructure	ID9, ID19, ID20, ID21, ID23, ID24, ID25, ID26, ID27, ID30, ID35, ID36, ID37
End-node (HW)	ID4, ID8, ID16, ID18, ID29, ID34
Network	ID1, ID17, ID22, ID33
Fog infrastructure	ID15

6. Discussion

The data presented in the previous section allowed us to answer the RQs that had guided the systematic mapping. In this section, we interpret the implications of the obtained results and provide a general discussion about the work performed.

First, we want to highlight that, despite the contemporary nature of the Edge AI topic, there is a surprisingly small amount of specific work about its assurance (we selected only 38 studies meeting the criteria). This reflects the immaturity of the Edge AI assurance domain and reinforces the relevance of this SMS, which shows how limited and scattered the current research efforts are (e.g., lack of consolidated practices, terminologies, and standards) and establishes a foundation for future studies. Assurance is a critical aspect for the adoption of Edge AI, ensuring that it performs as intended and complies with certain requirements (e.g., ethical, legal, and technical standards). Without assurance, the adoption of Edge AI in practice is hindered, especially in critical domains.

Additionally, we want to emphasize that among the 200 unique authors across the 38 selected studies, only 24 are affiliated with industry, representing 12% of the total number of authors. This relatively low proportion indicates that the perspective of industry is not yet strongly reflected in the current literature on Edge AI assurance. Consequently, some research contributions may lack alignment with real-world industrial challenges. This limitation highlights the need to strengthen academia–industry collaboration to develop assurance approaches that are both rigorous and applicable.

Accordingly, future research should engage industrial partners to refine and validate these methods in practical settings.

We also need to highlight that a significant number of retrieved papers that are, *a priori*, related to Edge AI assurance were not considered in the final selection. For example, we identified several papers using blockchain as an assurance method, operating under the assumption that the use of blockchain is secure by default given its decentralised nature [81]. We discarded them since the purpose of this paper is to analyse how Edge AI assurance is actively performed. Additionally, numerous papers about the application of AI for assurance purposes were found, such as employing machine learning for assurance at the edge, rather than for assuring AI at the edge. These were also discarded.

Second, regarding the dependability concerns that have been addressed (RQ1), most of the selected studies focused on assuring safety or security (21 studies). This was expected since safety and security are the most prominent topics in the assurance domain. However, we would like to highlight that research needs to also focus on other concerns to deem an Edge AI system dependable (e.g., reliability). Ideally, assurance should be provided for as many dependability concerns as possible and, preferably, in a multi-concern way. For example, it is not enough to address only safety, but security must also be considered as well because of the connectivity features of Edge AI. This aligns with the conclusion of Pekaric et al. [61], who stated that treating dependability concerns separately leads to gaps in comprehensive risk assessment. However, multi-concern assurance has not been properly investigated yet. We acknowledge the difficulties of such multi-concern assurance (e.g., technical difficulties, lack of mature standards or tools, lack of evaluation metrics), but addressing this gap is essential for the advancement of Edge AI. Future research should explore methods for capturing assurance interdependencies, balancing conflicting quality attributes, and establishing systematic frameworks for the combined assurance of safety, security, and other dependability concerns.

Third, with regards to the domain analysis of the results (RQ2), we observed that the industry domain has a leading position on Edge AI assurance research. A reason for this is that this domain is one of the most susceptible to incorporating Edge AI. Intelligent devices, such as robots, are used daily in industries. Ensuring their security, for example, is crucial for the safety of workers. This was outlined in Vyhmeister and Castane [75]. However, several studies do not consider a specific domain (11 studies), or they claim to be for a general technology (11 studies, e.g., IoT and CPS). *A priori*, offering 'general' assurance may appear beneficial due to its potential for reuse. However, each specific domain typically presents unique characteristics that must be considered accordingly (e.g., regulations for assuring a system for an airplane vs a system for a bank). This variability implies that assurance strategies for Edge AI systems must be contextually adapted to the domain-specific constraints and requirements. So, these studies will most probably need to be adjusted if one intends to put their proposals into practice in a specific context.

Fourth, the SMS revealed that, surprisingly, most of the studies (28 out of 38 studies) do not consider standards for Edge AI assurance (RQ3). This result is aligned with the fact that not many industrial authors were identified as described above (only 12% of the total amount of authors have industrial affiliations). In combination, both results show that the perspective of industry is insufficiently represented yet in the retrieved literature. Although we acknowledge that there is not a dedicated standard for Edge AI considering its own particularities (e.g., runtime verification), we expected that, at least, existing standards about dependability were more considered (e.g., safety standards). In critical systems, for example, the compliance with standards is crucial [1]. In addition, some of the standards used are standards dedicated to other aspects (e.g., connectivity, interoperability) that may or may not include dependability concerns (e.g., IEEE 802.1 TSN).

Fifth, regarding the types of assurance evidence that have been managed (RQ4), we want to highlight that all the selected studies provide more than one type of evidence. This strengthens the reliability of the presented assurance (i.e., the more evidence, the more confidence). It was expected that evidence related to the definition of the Edge AI architecture would be found (21 studies). The distributed architecture of Edge AI (see Section 2) is one of its main particularities and thus it seems logical to provide evidence about its assurance. We also found evidence related to runtime (e.g., runtime verification). This was not included in the evidence taxonomy used as reference [54] since it was not developed for Edge AI. For the context of this work, finding evidence related to runtime makes sense since, by definition, the context of Edge AI systems is very changing (e.g., new data coming sensors).

One surprising aspect is that little hazard analysis evidence was found (only two studies provided evidence for hazard specification). In general, hazard analysis has significant weight in assuring 'traditional' (non-Edge AI) systems. Therefore, it seems reasonable that it might need to be considered for Edge AI as well. For example, specifying how to reduce the likelihood of hazards and the consequences when a hazard cannot be eliminated might be relevant for critical Edge AI systems. In addition, no historical evidence was found. This seems plausible since Edge AI systems are relatively new and no dependability specification based on past assurance exist.

Sixth, when referring to the dependability justification techniques that have been used (RQ5), the most identified technique was *argumentation* (11 studies used it). This result is not surprising since it is also the most-widely identified technique in safety assurance [54]. However, although types of evidence were indeed provided (RQ4), most of the studies (27 studies) do not used any justification technique for this evidence. Without justification techniques, it is difficult to determine how the evidence is structured and presented in a suitable way for, for example, compliance management. Future work must focus on this aspect for a wider adoption of Edge AI.

Seventh, regarding the AI techniques that have been considered for assurance (RQ6), ML (including its types, deep learning and federated learning) are the ones that have been most commonly identified. This coincides with most of the related work on AI assurance in general, e.g., [70,73,74] (Section 3). *A priori*, this seems plausible since these algorithms are the most widely used in AI. However, given the widespread adoption of other techniques such as computer vision (e.g., robotics) or *Natural Language Processing* (in generative AI), future research must also focus on assuring them to enhance its trust and confidence.

Finally, we found that the end nodes (either their software or hardware) are the most considered edge-specific characteristic (RQ7). Initially, it seems reasonable to focus on assuring the end nodes, as they could be the most vulnerable and 'important' part of the edge infrastructure. However, we cannot neglect to assure other characteristics inherent to the edge that could also be critical for the proper functioning of the system (e.g., communication between nodes or with the cloud).

In combination with the results of RQ6, it can be observed that some studies discuss assuring AI algorithms at the end nodes (without specifying any details), while other studies go further and detail the assurance for the execution of an AI algorithm in a specific component within the final node. For example, this is the case of study ID4, which deals with how to assure an ML algorithm within a processor (with accelerators) at the end node.

7. Threats to validity

As any other empirical study, we face threats to validity: **selection bias, inaccuracy in data extraction and analysis, generalization, and reliability**. We report them below according to frequent threats specific to secondary studies [2], including mitigation actions.

First, to ensure that the mapping is as *complete* as possible and that no important literature is missing, we used a hybrid strategy, considering a search in Scopus plus a snowballing process with Google Scholar, in addition to an exploratory search. Scopus indexes publications of the most important conferences and journals on the topic (i.e., Edge AI assurance). In addition, by scanning the references of the retrieved studies (i.e., backward reference search) and their citations (i.e., forward search), we contributed to the completeness of the SMS. By following the snowballing strategy, we also mitigated threats related to the lack of a standard use of terminology, or lack of any relevant term in the search string, which is a common threat to validity in a pure string-search based methodology.

We acknowledge that certain forms of grey literature may not have been systematically identified or included in the hybrid strategy (e.g., internal government reports, training material, industry manuals). However, the inclusion of these sources could present challenges related to reproducibility, limited accessibility, and lack of formal quality control. This constitutes a limitation of our work and highlights a future direction aimed at enriching the understanding of Edge AI assurance. Additionally, given that the study selection process involved human judgment (particularly when reading titles and abstracts) there is an inherent risk of subjectivity and bias in both the inclusion and exclusion of studies.

Second, *data extraction and analysis* were carried out mainly by the first author. This might comprise subjective decisions and interpretations when extracting information from reading the full text of the primary studies. To mitigate this risk, a rigorous extraction process was applied based on the guidelines of SMS [48,62]. In addition, the co-authors continuously checked the work of the first author resolving disagreements through discussion.

Third, although the analysis presented in this SMS is based on a rigorous selection process from over 3000 initially identified studies, the resulting set of 38 primary papers reflects the novelty and limited maturity of the Edge AI assurance field. While this small number enhances the visibility of existing research, it also constrains the extent to which *generalizations* can be drawn at this stage. As such, the identified findings should be interpreted as indicative rather than definitive, representing a snapshot of the current literature rather than a complete picture.

Finally, we ensure *reliability* as the search process can be replicated by other researchers because we have documented the steps performed and shared the intermediate results. However, since the data extraction process also considers subjective factors (e.g., interpretation of the full text of the studies), we cannot guarantee that other researchers will obtain exactly the same results as presented in this work.

8. Conclusions and future work

Novel Edge AI applications pose new challenges for system developers, e.g., regarding physical limitations and response time. Assurance of these applications is not an easy task. Our work aims to provide a fundamental understanding of Edge AI assurance. For this purpose, a systematic mapping study was conducted. We retrieved a total of 3113 studies of which only 38 were identified as primary studies for providing such an understanding. After analysing these studies, we found that ten dependability concerns have already been addressed (e.g., safety and security), although some others have not been considered yet (e.g., reliability) nor their relationships (multi-concern assurance). In

addition, we studied the different application domains in which Edge AI assurance has been applied, finding that the industry domain is the prominent one. We also have identified that not many standards have been considered yet, which are of real importance specially for critical systems. In the SMS, although we found 27 different types of assurance evidence, we identified few techniques to justify dependability (e.g., argumentation). Finally, regarding AI- and edge-specific characteristics, we found that machine learning algorithms (including their subtypes) running on the end nodes are the most widely considered.

This work is, to our knowledge, the only existing review on the topic of Edge AI assurance. The results provide useful insights for both researchers and practitioners. From a research perspective, new research gaps have been identified (e.g., hazard analysis for Edge AI). As for practitioners, the results provide a concrete reference for understanding the main aspects of Edge AI assurance (e.g., what needs to be considered, such as architecture specifications and runtime verification). However, we emphasize that the SMS is focused on literature. Subsequently, no strong conclusions can be drawn on Edge AI assurance in practice. Analysing practical usefulness and industrial adoption requires further studies on the current state of practice. In addition, although this SMS aimed to provide a general overview, future work should explore specific adaptations of Edge AI assurance practices (e.g., depending on the domain), including the tailoring of dependability concerns, standards, and evidence types to distinct industrial contexts.

The SMS is part of an on-going research effort aimed at improving Edge AI assurance. The aggregation of the knowledge extracted in this work will result in an Edge AI assurance framework: a specification of needs that might have to be considered to provide adequately justified confidence that Edge AI is dependable. The framework is intended to be generic so that it can be used in different scenarios. However, since certain dependability attributes (e.g., safety) are highly dependent on the specific Edge AI application and its usage context (e.g., drone versus plane), the framework will need to be adapted to the specific characteristics of this context. For future work, we plan to apply it in the context of the REBECCA project. In particular, we will apply it for the four project use cases: underwater robots, unmanned aerial vehicles, AI-powered fridges, and industrial equipment inspection. They are concrete examples of an Edge AI system and of Edge AI applications. The outcome from applying the assurance framework will contribute to providing evidence of the degree of confidence in the dependability of the REBECCA's results in a specific context.

Acknowledgement. The work leading to this paper has received funding from the REBECCA (HORIZON-KDT ref. 101097224; MCIN/AEI ref. PCI2022-135043-2; NextGen.EU/PRTR), AETERNAL (MCIN/AEI ref. PID2023-149753OB-C21; ERDF), FDT4S (SBPLY/24/180225/000020, ERDF), "Paradigmas de interacción para la nueva era de resiliencia digital" (UCLM ref. 2022-GRIN-34436; ERDF), and "Una propuesta integral para el desarrollo independiente de dominio de gemelos digitales" (UCLM 2025-GRIN-38441; ERDF) projects.

Appendices

A. Primary studies (sorted by order of appearance during the search)

ID1. Chahed, H., Usman, M., Chatterjee, A., Bayram, F., Chaudhary, R., Brunstrom, A., Taheri, J., Ahmed, B.S., & Kassler, A. (2023). AIDA—A holistic AI-driven networking and processing framework for industrial IoT applications. *Internet of Things*, *22*, 100805.

ID2. Hawkins, R., Picardi, C., Donnell, L., & Ireland, M. (2023). Creating a safety assurance case for a machine learned satellite-based wildfire detection and alert system. *Journal of Intelligent & Robotic Systems*, *108*(3), 47.

ID3. Nagy, S. J., Szabó, R., Vajda, M. L., & Vörös, A. (2021). Demonstrator for dependable edge-based cyber-physical systems. In *2021 10th Latin-American Symposium on Dependable Computing (LADC)* (pp. 1-8). IEEE.

ID4. Gomony, M. D., Gebregiorgis, A., Fieback, M., Geilen, M., Stuijk, S., Richter-Brockmann, J., Bisnoi, R., Argo, S., Arche Andradas, L., Güneysu, T., Taouil, M., Corporaal, H., & Hamdioui, S. (2023). Dependability of Future Edge-AI Processors: Pandora's Box. In *2023 IEEE European Test Symposium (ETS)* (pp. 1-6). IEEE.

ID5. Hung, Y. W., Chen, Y. C., Lo, C., So, A. G., & Chang, S. C. (2021). Dynamic workload allocation for edge computing. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, *29*(3), 519-529.

ID6. Rehman, A., Awan, K. A., Ud Din, I., Almogren, A., & Alabdulkareem, M. (2023). FogTrust: fog-integrated multi-leveled trust management mechanism for internet of things. *Technologies*, *11*(1), 27.

ID7. Bouchekir, R., Guzman, M., Cook, A., Haindl, J., & Woolnough, R. (2023). Formal verification for safe ai-based flight planning for uavs. In *2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)* (pp. 275-282). IEEE.

ID8. Ibtesam, M., Solangi, U. S., Kim, J., Ansari, M. A., & Park, S. (2021). Highly Efficient Test Architecture for Low-Power AI Accelerators. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *41*(8), 2728-2738.

ID9. Duarte, L. O., & Prestes, J. D. L. (2021). IoT solution information security certification conceptual framework. In *XVII Brazilian Symposium on Information Systems*.

ID10. Drusinsky, D., Litton, M., & Michael, J. B. (2022). Lightweight verification and validation of cyberphysical systems using machine-learned correctness properties. *Computer*, *55*(2), 102-108.

ID11. Jaradat, O., Sljivo, I., Hawkins, R. D., & Habli, I. (2020). Modular safety cases for the assurance of industry 4.0. In *Safety-Critical Systems Symposium*.

ID12. Kim, M. J., Hong, S. P., Kang, M., & Seo, J. (2021). Performance comparison of posenet models on an aiot edge device. *Intelligent Automation & Soft Computing*, *30*(3), 743-753.

ID13. Anisetti, M., Ardagna, C. A., Bena, N., & Damiani, E. (2023). Rethinking certification for trustworthy machine-learning-based applications. *IEEE Internet Computing*, *27*(6), 22-28.

ID14. Bacciu, D., Carta, A., Gallicchio, C., & Schmittner, C. (2023). Safety and robustness for deep neural networks: An automotive use case. In *International Conference on Computer Safety, Reliability, and Security* (pp. 95-107). Cham: Springer Nature Switzerland.

ID15. Desai, N., & Punnekkat, S. (2019). Safety of fog-based industrial automation systems. In *Proceedings of the Workshop on Fog Computing and the IoT* (pp. 6-10).

ID16. Bena, N., Bondaruc, R., & Polimeno, A. (2022). Security Assurance in Modern IoT Systems. In *2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)* (pp. 1-5).

ID17. Sedar, R., Vázquez-Gallego, F., Casellas, R., Vilalta, R., Muñoz, R., Silva, R., Dizambourg, L., Fernández Barciela, A.E., Vilajosana, X., Kanti Datta, S., Härri, J., & Alonso-Zarate, J. (2021). Standards-compliant multi-protocol on-board unit for the evaluation of connected and automated mobility services in multi-vendor environments. *Sensors*, *21*(6), 2090.

ID18. Ramírez-Bárcenas, A., Portela-García, M., García-Valderas, M., & López-Ongil, C. (2020). System dependability in edge computing wearable devices. In *2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS)* (pp. 1-6).

ID19. Bacciu, D., Akarmazyan, S., Armengaud, E., Bacco, M., Bravos, G., Calandra, C., Carlini, E., Carta, A., Cassarà, P., Coppola, M., Davalas, C., Dazzi, P., Degennaro, M.C., Di Sarli, D., Dobaj, J., Callicchio, C., Girbal, S., Gotta, A., Groppo, R., Lomonaco, V., Macher, G., Mazzei, D., Mencagli, G., Michail, D., Micheli, A., Peroglio, R., Petroni, S., Potenza, R., Pourdanesh, F., Sadianos, C., Tserpes, K., Tagliabó, F., Valtl, J., Varlamis, I., & Veledar, O. (2021). Teaching-trustworthy autonomous cyber-physical applications through human-centred intelligence. In *2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS)* (pp. 1-6).

ID20. Hendriks, T., Akesson, B., Voeten, J., Hendriks, M., Parada, J. C., García-Gordillo, M., Sáez, S., & Valls, J. J. (2023). Thirteen concepts to play it safe with the cloud. In *2023 IEEE International Systems Conference (SysCon)* (pp. 1-7).

ID21. Javed, M. A., Muram, F. U., Hansson, H., Punnekkat, S., & Thane, H. (2021). Towards dynamic safety assurance for Industry 4.0. *Journal of Systems Architecture*, *114*, 101914.

ID22. Qureshi, K. N., Iftikhar, A., Bhatti, S. N., Piccialli, F., Giampaolo, F., & Jeon, G. (2020). Trust management and evaluation for edge intelligence in the Internet of Things. *Engineering Applications of Artificial Intelligence*, *94*, 103756.

ID23. Pandey, A., Calyam, P., Debroy, S., Wang, S., & Alarcon, M. L. (2021). Vectrust: Trusted resource allocation in volunteer edge-cloud computing workflows. In *Proceedings of the 14th IEEE/ACM international conference on utility and cloud computing* (pp. 1-10).

ID24. Abdel-Basset, M., Moustafa, N., & Hawash, H. (2022). Privacy-preserved cyberattack detection in Industrial Edge of Things (IEoT): A blockchain-orchestrated federated learning approach. *IEEE Transactions on Industrial Informatics*, *18*(11), 7920-7934.

ID25. Jin, Y., Huang, B., Yan, Y., Huan, Y., Xu, J., Li, S., Gope, P., Xu, L.D., Zou, Z., & Zheng, L. (2022). Edge-based collaborative training system for artificial intelligence-of-things. *IEEE Transactions on Industrial Informatics*, *18*(10), 7162-7173.

ID26. Drusinsky, D., Michael, J. B., & Litton, M. (2022). Machine-learned specifications for the verification and validation of autonomous cyberphysical systems. In *2022 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)* (pp. 333-341).

ID27. Anisetti, M., Ardagna, C. A., Bena, N., Damiani, E., & Panero, P. G. (2023). Managing ML-Based Application Non-Functional Behavior: A Multi-Model Approach. *CoRR abs/2311.12686 arXiv:2311.12686.*

ID28. Awan, K. A., Ud Din, I., Almogren, A., Khattak, H. A., & Rodrigues, J. J. (2022). EdgeTrust: A lightweight data-centric trust management approach for IoT-based healthcare 4.0. *Electronics*, *12*(1), 140.

ID29. Silva Neto, A. V., Silva, H. L., Camargo Jr, J. B., Almeida Jr., J. R., & Cugnasca, P. S. (2023). Design and Assurance of Safety-Critical Systems with Artificial Intelligence in FPGAs: The Safety ArtISt Method and a Case Study of an FPGA-Based Autonomous Vehicle Braking Control System. *Electronics*, *12*(24), 4903.

ID30. Arachchige, P. C. M., Bertok, P., Khalil, I., Liu, D., Camtepe, S., & Atiquzzaman, M. (2020). A trustworthy privacy preserving framework for machine learning in industrial IoT systems. *IEEE Transactions on Industrial Informatics*, *16*(9), 6092-6102.

ID31. Hathout, B., Shepherd, P., Dagiuklas, T., Nagaty, K., Hamdy, A., & Rodriguez, J. (2024). Adaptive Trust Management for Data Poisoning Attacks in MEC-based FL Infrastructures. *IEEE Open Journal of the Communications Society 6*, 3140-3160.

ID32. Kotilainen, P., Mäkitalo, N., Systä, K., Mehraj, A., Waseem, M., Mikkonen, T., & Murillo, J. M. (2025). Allocating distributed AI/ML applications to cloud–edge continuum based on privacy, regulatory, and ethical constraints. *Journal of Systems and Software*, *222*, 112333.

ID33. Berto, F., Ardagna, C. A., Banzi, M., & Anisetti, M. (2024). Assurance in Advanced 5G Edge Continuum. *IEEE Access 12,* 178659-178671.

ID34. Leveugle, R., Benabdenbi, M., Al Kaf, A., & Noizette, L. (2024). Combining Acceleration and Approximation in Dependable Edge AI: Optimization Methodology and Tools Applied to a Case Study. In *IEEE International Conference on Design, Test and Technology of Integrated Systems (DTTIS'24)* (pp. 1-6). IEEE.

ID35. Sedlak, B., Pujol, V. C., Donta, P. K., & Dustdar, S. (2024). Equilibrium in the computing continuum through active inference. *Future Generation Computer Systems*, *160*, 92-108.

ID36. Withana, S., & Plale, B. (2024). Patra ModelCards: AI/ML Accountability in the Edge-Cloud Continuum. In *2024 IEEE 20th International Conference on e-Science (e-Science)* (pp. 1-10). IEEE.

ID37. Anisetti, M., Ardagna, C. A., & Bena, N. (2023). Continuous certification of non-functional properties across system changes. In *International Conference on Service-Oriented Computing (ICSOC'18)* (pp. 3-18). Springer.

ID38. Doku, R., & Rawat, D. B. (2021). Mitigating data poisoning attacks on a federated learning-edge computing network. In *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)* (pp. 1-6). IEEE.

B. Examples of data extracted from the primary studies

ID	Found in	Publication year	Concern (RQ1)	Domain (RQ2)	Standard (RQ3)	Evidences (RQ4)	Dependability justification techniques (RQ5)	AI techniques (RQ6)	Edge-specific characteristics (RQ7)
1	Scopus	2023	Trustworthiness	Industry 4.0	IEEE 802.1 TSN	Architecture specification, Runtime verification (o monitoring?) results, testing results (reliability or robustness testing)	0	Machine learning	Network
2	Exploratory	2023	Safety	Space	No standard	requirements specification (safety, data), Review results (data, ML functionality), Testing results (Functional testing results, Test cases specification (verification data)) , Activity records (model development	Argumentation (Safety assurance case (GSN))	Machine learning	End node(SW, ML included aboard)
3	Scopus	2021	Dependability	Smart city	No standard	Architecture specification	0	Deep learning	End node (SW)
4	Exploratory	2023	Dependability	No domain	No standard	Architecture specification, Testing results (qué tipo?), Runtime verification results	0	AI models/algorithms	End node (HW, Edge-AI processor)
5	Scopus	2021	Accuracy	No domain	No standard	Runtime verification results	0	Deep learning	End node (SW, Predictions (deep neural) done at the edge are trustworthy or not)
6	Scopus	2023	Security	No domain	No standard	Simulation results	0	AI models/algorithms	End node (SW, detect and eliminate malicious edge nodes)
7	Exploratory	2023	Safety	UAV	DO-333	Theorem proving results, Model checking results, Requirements specification, Review results	Argumentation (GSN)	Pathfinding	End node (planning component of the UAVs)
8	Scopus	2022	Safety	Automotive	ISO 26262	Tool support specification, Testing results	0	AI models/algorithms	End node (HW, AI Accelerators at the edge)
9	Exploratory	2021	Security	IoT	No standard	V&V results, Reused component specification, Activity records, Source code, Requirements specification, V&V plan, Object code	0	AI models/algorithms	Edge-AI infrastructure (smart IoT solutions)
10	Exploratory	2022	Safety	CPS	No standard	Theorem proving results, Model checking results, Runtime verification, Requirements specification, Testing results (regression, unit), Test cases specification	Argumentation (assurance argument) - structured text	Machine learning	End node (SW, CPS)
11	Exploratory	2020	Safety	Industry 4.0	No standard	Review results, Requirements specification (assurance contract, safety requirements), Hazard specification, Risk analysis results (Safety analysis results (HARA))	Argumentation (Modular safety cases) GSN	AI models/algorithms	End node (SW)
12	Scopus	2021	Performance	Video capturing/stream	oneM2M	Architecture specification	0	Deep learning	End node (SW)
13	Exploratory	2023	Security	No domain	No standard	Testing results (robustness, operational ("behavior of the ML model in operation")), Runtime verification results, Test cases specification, Architecture specification, Automated static analysis results, Inspection results, "(training) data, (training) process, and the ML model itself" --> Requirements specification, Review results, Activity records, Tool support specification, Design specification	0	Machine learning	End node (SW)

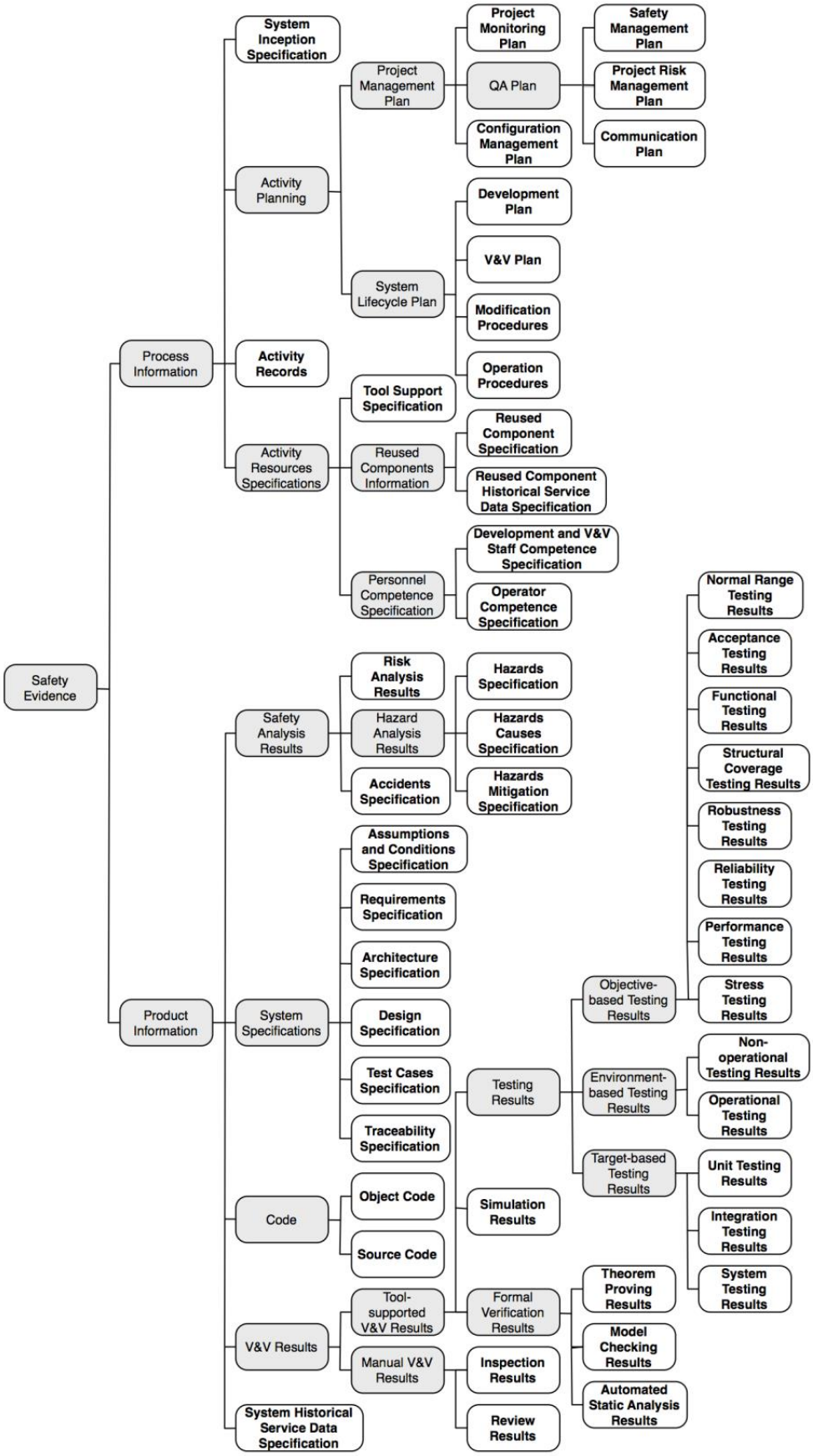
C. Evidence types

This appendix provides the definition of the evidence types found (alphabetically ordered) to facilitate their understanding. Most of the definitions are extracted from [54]. However, since it was developed for safety in critical systems in general, we have included 'runtime verification results' as a new type of evidence to reflect the current needs of Edge AI assurance. For illustration purposes, each evidence type is exemplified with content from the primary studies. We include techniques, artefacts, or information that can be considered as or used to for the identification of the evidence type. Finally, a screenshot of the taxonomy in [54] is also presented.

- **Activity records:** Specification of the work performed to execute the activity planning of a system. Example: model development logs (artefact) in ID2.
- **Architecture specification:** Description of the fundamental organisation of a system, embodied in its components, their relationships to each other and to the environment, and the principles guiding its design and evolution. Examples: architecture diagram in ID1, software architectures in ID3 and ID17, architectures of PoseNets in ID12, and architecture of the edge-cloud continuum in ID27 (artefacts).
- **Automated static analysis results:** Results from an automatic process for evaluating a system based on its form, structure, content, or documentation. Example: integrity analysis (technique) in ID13.
- **Design specification:** Specification of the components, interfaces, and other internal characteristics of a system or component. Examples: flow diagram in ID12, threat and system model in ID24, and flowchart of the dataset generator in ID29 (artefacts).
- **Development plan:** Description of how a system will be built. It includes information about the requirements, design, and implementation (coding and/or integration) phases. Example: developing activities to be performed (information) in ID29.
- **Hazard specification:** Specification of the conditions in a system that can become a unique, potential accident. Examples: hazard analysis (technique) in ID11 and ID21.
- **Inspection results:** Results from the visual examination of system lifecycle work products of a system to detect errors, violations of development standards, and other problems. Example: inspection of checkpoints (technique) in ID13.
- **Management plan:** Description of the coordinated, comprehensive set of processes designed to direct and control resources to optimally manage the operational aspect of an organisation. Examples: dependability management plan in ID18, data management plan in ID32, certification management plan in ID37 (artefacts).
- **Model checking results:** Results from the verification of the conformance of a system to a given specification by providing a formal guarantee. The system under verification is modelled as a state transition system, and the specifications are expressed as temporal logic formulae that express constraints over the system dynamics. Example: the evidence for the safety cases are obtained through model checking (technique) in ID21.
- **Modification procedure:** Description of the instructions as to what to do when performing a modification in a critical system to make corrections, enhancements, or adaptations to the validated system, ensuring that the required requirement is sustained. Example: change detection system (artefact) in ID37.
- **Object code:** Computer instructions and data definitions in a form output by an assembler or compiler. Example: digital signed result of a certification contract execution in ID9.

- **Operator competence specification:** Specification of the skills or knowledge that the parties involved in the operation procedures need to carry out the activities assigned to them. Example: operator competence description (information) in ID26.
- **Project risk management plan:** Description of the activity regarding the development and documentation of an organised and comprehensive strategy for identifying project risks. It includes establishing methods for mitigating and tracking risk. Example: project risk identification (technique) in ID32.
- **Requirements specification:** Specification of the external conditions and capabilities that a system must meet and possess, respectively, to (1) allow a user to solve a problem or achieve an objective, or (2) satisfy a contract, standard, specification, or other formally imposed documents. Examples: software requirements specification (artefact) in ID16 and ID31.
- **Reused component specification:** Specification of the characteristics of an existing system that is (re-) used to make up a system. Example: reused component specification (artefact) in ID9.
- **Review results:** Description of a process or meeting during which a system lifecycle work product or set of works products is presented to some interested party for comment or approval. Example: data review (artefact) in ID2.
- **Risk analysis results:** Specification of the expected amount of danger when an identified hazard will be activated and thus become an accident in a system. Examples: hazard analysis and risk assessment in ID11 and risk evaluation in ID23 (techniques).
- **Runtime verification results:** Results from the analysis of a system's behaviour during its operation (after deployment) to determine if it complies with predefined properties or assurance requirements. Examples: runtime verification (technique) in ID5, ID20, ID25, ID30, and ID38.
- **Simulation results:** Results from the verification of a system by creating a model that behaves or operates like the system when provided with a set of controlled inputs. Examples: experimental simulation (technique) in ID6 and ID28.
- **Source code:** Computer instructions and data definitions expressed in a form suitable for input to an assembler, compiler, or other translator. Example: deployment code (artefact) in ID36.
- **System inception specification:** Specification of initial details about the characteristics of a system and how it will be created. Example: initial system design (artefact) in ID19.
- **Test cases specification:** Description of the purpose, inputs, expected outcomes, and steps for executing a specific test case, along with criteria for success. Examples: regression testing (technique) in ID10, robustness and stress testing (technique) in ID14, service level objectives testing (technique) in ID35.
- **Testing results:** Documented outcomes of test executions, comparing the actual system or software behaviour with expected results. Examples: quality test results in ID4, power consumption testing results in ID8, and reliability testing results in ID15 (artefacts).
- **Theorem proving results:** Formal, mathematically verified results to demonstrate certain properties or behaviours of a system. Example: lemma results with the position of the aircrafts (artefact) in ID7.
- **Tool support specification:** Capabilities, integration requirements, and operational constraints of tools used to support assurance processes. Example: specification of the fault injection tool (artefact) in ID34.

- **V&V results:** The documented outcomes of the Verification and Validation (V&V) process. Verification involves checking that the system is built correctly according to design specifications, while validation ensures that the system fulfils its intended purpose and meets user needs. Example: verification results (artefact) in ID33.
- **V&V plan:** Structured processes and methodologies to ensure that a system meets its specified requirements (verification) and fulfils its intended purpose (validation). Example: validation plan (technique) in ID9.


Safety Evidence
Process Information
System Inception Specification
Activity Planning
Project Management Plan
Project Monitoring Plan
QA Plan
Safety Management Plan
Project Risk Management Plan
Configuration Management Plan
Communication Plan
System Lifecycle Plan
Development Plan
V&V Plan
Modification Procedures
Operation Procedures
Activity Records
Activity Resources Specifications
Tool Support Specification
Reused Components Information
Reused Component Specification
Reused Component Historical Service Data Specification
Personnel Competence Specification
Development and V&V Staff Competence Specification
Operator Competence Specification
Product Information
Safety Analysis Results
Risk Analysis Results
Hazard Analysis Results
Accidents Specification
Hazards Specification
Hazards Causes Specification
Hazards Mitigation Specification
System Specifications
Assumptions and Conditions Specification
Requirements Specification
Architecture Specification
Design Specification
Test Cases Specification
Traceability Specification
Code
Object Code
Source Code
V&V Results
Tool-supported V&V Results
Manual V&V Results
System Historical Service Data Specification
Testing Results
Simulation Results
Formal Verification Results
Inspection Results
Review Results
Objective-based Testing Results
Environment-based Testing Results
Target-based Testing Results
Normal Range Testing Results
Acceptance Testing Results
Functional Testing Results
Structural Coverage Testing Results
Robustness Testing Results
Reliability Testing Results
Performance Testing Results
Stress Testing Results
Non-operational Testing Results
Operational Testing Results
Unit Testing Results
Integration Testing Results
System Testing Results
Theorem Proving Results
Model Checking Results
Automated Static Analysis Results

Declaration of generative AI and AI-assisted technologies in the writing process.

During the preparation of this work the authors used Copilot (https://copilot.microsoft.com) in some parts of the text in order to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Adler, R., Akram, M. N., Bauer, P., Feth, P., Gerber, P., Jedlitschka, A., Jöckel, L., Kläs, M., & Schneider, D. (2019). Hardening of Artificial Neural Networks for Use in Safety-Critical Applications--A Mapping Study. *arXiv preprint arXiv:1909.03036*

2. Ampatzoglou, A., Bibi, S., Avgeriou, P., Verbeek, M., & Chatzigeorgiou, A. (2019). Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. *Information and software technology*, *106*, 201-230.

3. American National Standards Institute/Industrial Truck Safety Development Foundation, Safety standard for driverless, automatic guided industrial vehicles and automated functions of manned industrial vehicles, (2019), http://www.itsdf.org.

4. Amiri, Z., Heidari, A., Navimipour, N. J., & Unal, M. (2023). Resilient and dependability management in distributed environments: A systematic and comprehensive literature review. *Cluster Computing*, *26*(2), 1565-1600.

5. Ashouri, M., Davidsson, P., & Spalazzese, R. (2021). Quality attributes in edge computing for the Internet of Things: A systematic mapping study. *Internet of Things*, *13*, 100346.

6. Avizienis, A., Laprie, J. C., Randell, B., & Landwehr, C. (2004). Basic concepts and taxonomy of dependable and secure computing. *IEEE transactions on dependable and secure computing*, *1*(1), 11-33.

7. Ayora, C., Torres, V., Weber, B., Reichert, M., & Pelechano, V. (2015). VIVACE: A framework for the systematic evaluation of variability support in process-aware information systems. *Information and Software Technology*, *57*, 248-276.

8. Ayora, C., García, A. S., & de la Vara, J. L. (2024). Edge-AI Assurance in the REBECCA Project. In *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (pp. 594-597).

9. Ayora, C., García A.S., de la Vara, J. L. (2025) Data extraction SMS Edge-AI assurance. https://doi.org/10.5281/zenodo.16326422

10. Babeshko, I., & Di Giandomenico, F. (2023). Safety and cybersecurity assessment techniques for critical industries: A mapping study. *IEEE Access*, *11*, 83781-83793.

11. Bardis, N. G., Doukas, N., Kharchenko, V., Sklyar, V., & Yaremchuk, S. (2022). Approaches and Techniques to Improve IoT Dependability. In *Dependable IoT for Human and Industry* (pp. 307-328). River Publishers.

12. Canpolat Şahin, M., & Kolukısa Tarhan, A. (2025). Evaluation and Selection of Hardware and AI Models for Edge Applications: A Method and A Case Study on UAVs. *Applied Sciences*, *15*(3), 1026.

13. Chance, G., Abeywickrama, D. B., LeClair, B., Kerr, O., & Eder, K. (2023). Assessing trustworthiness of autonomous systems. *arXiv preprint arXiv:2305.03411*.

14. Chandramouli, R. (2013). Security Assurance Requirements for Hypervisor Deployment Features. In *Seventh International Conference on Digital Society*.

15. Chinnasamy, P., Rojaramani, D., Praveena, V., SV, A. J., & Bensujin, B. (2021). Data Security and Privacy Requirements in Edge Computing: A Systemic Review. *Cases on Edge Computing and Analytics*, 171-187.

16. Costal, D., Farré, C., Franch, X., & Quer, C. (2021). How tertiary studies perform quality assessment of secondary studies in software engineering. *arXiv preprint arXiv:2110.03820*.

17. Dai, W., Nishi, H., Vyatkin, V., Huang, V., Shi, Y., & Guan, X. (2019). Industrial edge computing: Enabling embedded intelligence. *IEEE Industrial Electronics Magazine*, *13*(4), 48-56.

18. de la Vara, J. L., García, A. S., Valero, J., & Ayora, C. (2022). Model-based assurance evidence management for safety–critical systems. *Software and Systems Modeling*, *21*(6), 2329-2365.

19. de La Vara, J. L., Gallina, B., Fernández-Caballero, A., Molina, J. P., García, A. S., & Ayora, C. (2023). Assurance of software-intensive medical devices: What about mental harm?. In *2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S)* (pp. 168-172). IEEE.

20. de la Vara, J. L., Ruiz, A., & Blondelle, G. (2021). Assurance and certification of cyber–physical systems: The AMASS open source ecosystem. *Journal of systems and software*, *171*, 110812.

21. de la Vara, J. L., Ruiz, A., Attwood, K., Espinoza, H., Panesar-Walawege, R. K., López, Á., del Río, I., & Kelly, T. (2016). Model-based specification of safety compliance needs for critical systems: A holistic generic metamodel. *Information and software technology*, *72*, 16-30.

22. ECS. (2025). Strategic Research and Innovation Agenda 2025.

23. ETSI. (2021). Cooperative Intelligent Transport Systems (C-ITS). European Telecommunications Standards Institute.

24. EU AI Act (Regulation (EU) 2024/1689) https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

25. EU General Data Protection Regulation (Regulation (EU) 2016/679) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679

26. Gallina, B., Montecchi, L., De Oliveira, A. L., & Bressan, L. (2022). Multiconcern, dependability-centered assurance via a qualitative and quantitative coanalysis. *IEEE Software*, *39*(4), 39-47.

27. Garfinkel, S., Guttman, B., Near, J., Dajani, A. N., & Singer, P. (2023). De-Identifying Government Datasets: Techniques and Governance. National Institute of Standards and Technology, NIST Special Publication 800-188.

28. Gyllenhammar, M., de Campos, G. R., & Törngren, M. (2025). The Road to Safe Automated Driving Systems: A Review of Methods Providing Safety Evidence. *IEEE Transactions on Intelligent Transportation Systems*.

29. Hawkins, R., Paterson, C., Picardi, C., Jia, Y., Calinescu, R., & Habli, I. (2021). Guidance on the assurance of machine learning in autonomous systems (AMLAS). *arXiv preprint arXiv:2102.01564*.

30. Himeur, Y., Sayed, A. N., Alsalemi, A., Bensaali, F., & Amira, A. (2024). Edge AI for Internet of Energy: Challenges and perspectives. *Internet of Things*, *25*, 101035.

31. Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, *15*(9), 1277-1288.

32. IEC. (2011). IEC 61508 - Functional safety of electrical/electronic/programmable electronic safety-related systems, 2nd ed.

33. IEEE 802.1 Working Group, et al., IEC/IEEE 60802 TSN profile for industrial automation, (2021).

34. IEEE P2802 - Standard for the Performance and Safety Evaluation of Artificial Intelligence Based Medical Device, (2022).

35. IEEE Standard for Wireless Personal Area Networks (WPANs), IEEE Std 802.15, (2003).

36. IEEE Standard P2805.3 Cloud-Edge Collaboration Protocols for Machine Learning, (2019).

37. IEEE 1935-2023: IEEE Standard for Edge/Fog Manageability and Orchestration https://standards.ieee.org/ieee/1935/7181/ (2023).

38. Iskandar, U. A. S., Diah, N. M., & Ismail, M. (2020). Identifying artificial intelligence Pathfinding algorithms for Platformer games. In *2020 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)* (pp. 74-80). IEEE.

39. ISO/IEC 27001:2013. Information technology—Security techniques—Information security management systems—Requirements, (2013).

40. ISO/IEC 26262:2018. Road vehicles – Functional safety, (2018).

41. ISO/IEC TR 24028:2020. Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence, (2020).

42. ISO/IEC TR 5469:2024. Artificial intelligence — Functional safety and AI systems, (2024).

43. ISO/IEC 27031:2025. Cybersecurity — Information and communication technology readiness for business continuity, (2025).

44. ITU-T Y.3000-series http://demo.ifgict.org/wp-content/uploads/2024/08/ITU-Artificial-intelligence-standardization-roadmap.pdf

45. Khan, W. Z., Ahmed, E., Hakak, S., Yaqoob, I., & Ahmed, A. (2019). Edge computing: A survey. *Future Generation Computer Systems*, *97*, 219-235.

46. Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O. P., Turner, M., Niazi, M., & Linkman, S. (2010). Systematic literature reviews in software engineering–a tertiary study. *Information and software technology*, *52*(8), 792-805.

47. Kitchenham, B. A., Budgen, D., & Brereton, P. (2015). Evidence-based software engineering and systematic reviews. Chapman & Hall/CRC.

48. Kitchenham, B.A. (2007) Guidelines for Performing Systematic Literature Reviews in Software Engineering, Version 2.3. Keele University and University of Durham, EBSE Technical Report.

49. Kropatschek, S., Hollerer, S., Hoffman, D., Winkler, D., Lüder, A., Sauter, T., Kastner, W., & Biffl, S. (2023). Combining models for safety and security concerns in automating digital production. In *2023 IEEE 21st International Conference on Industrial Informatics (INDIN)* (pp. 1-8). IEEE.

50. Mansourov, N., & Campara, D. (2010). System assurance: beyond detecting vulnerabilities. Elsevier.

51. McDermid, J. A., Jia, Y., & Habli, I. (2019). Towards a framework for safety assurance of autonomous systems. In *Artificial Intelligence Safety 2019* (pp. 1-7). CEUR Workshop Proceedings.

52. Mohseni, S., Wang, H., Xiao, C., Yu, Z., Wang, Z., & Yadawa, J. (2022). Taxonomy of machine learning safety: A survey and primer. *ACM Computing Surveys*, *55*(8), 1-38.

53. Mourão, E., Pimentel, J. F., Murta, L., Kalinowski, M., Mendes, E., & Wohlin, C. (2020). On the performance of hybrid search strategies for systematic literature reviews in software engineering. *Information and software technology*, *123*, 106294.

54. Nair, S., De La Vara, J. L., Sabetzadeh, M., & Briand, L. (2014). An extended systematic literature review on provision of evidence for safety certification. *Information and Software Technology*, *56*(7), 689-717.

55. Nair, S., de la Vara, J.L., Sabetzadeh, M., Falessi, D. (2015): Evidence Management for Compliance of Critical Systems with Safety Standards: A Survey on the State of Practice. *Information and Software Technology* 60: 1-15.

56. Nascimento, A. M., Vismari, L. F., Molina, C. B. S. T., Cugnasca, P. S., Camargo, J. B., Almeida Jr., J. R., Inam, R., Fersman, E. Marquezini, V. & Hata, A. Y. (2019). A systematic literature review about the impact of artificial intelligence on autonomous vehicle safety. *IEEE Transactions on Intelligent Transportation Systems*, *21*(12), 4928-4946.

57. Neto, A. V. S., Camargo, J. B., Almeida, J. R., & Cugnasca, P. S. (2022). Safety assurance of artificial intelligence-based systems: A systematic literature review on the state of the art and guidelines for future work. *IEEE Access*, *10*, 130733-130770.

58. Oliveira, F., Costa, D. G., Assis, F., & Silva, I. (2024). Internet of Intelligent Things: A convergence of embedded systems, edge computing and machine learning. *Internet of Things*, 101153.

59. Omoniwa, B., Hussain, R., Javed, M. A., Bouk, S. H., & Malik, S. A. (2018). Fog/edge computing-based IoT (FECIoT): Architecture, applications, and research issues. *IEEE Internet of Things Journal*, *6*(3), 4118-4149.

60. oneM2M, Service layer core protocol specification, Release 2 TS-0004 v2.27.0, (2020).

61. Pekaric, I., Groner, R., Witte, T., Adigun, J. G., Raschke, A., Felderer, M., & Tichy, M. (2023). A systematic review on security and safety of self-adaptive systems. *Journal of Systems and Software*, *203*, 111716.

62. Petersen, K., Feldt R., Mujtaba, S., & Mattson, M. (2008): Systematic Mapping Studies in Software Engineering. In *12th international conference on evaluation and assessment in software engineering (EASE'08)*. DOI: 10.14236/ewic/EASE2008.8.

63. Rajabli, N., Flammini, F., Nardone, R., & Vittorini, V. (2020). Software verification and validation of safe autonomous cars: A systematic literature review. *IEEE Access*, *9*, 4797-4819.

64. RTCA DO-333, Formal Methods Supplement to DO-178C and DO-278A, (Washington, DC: RTCA, Inc., 2011).

65. Sánchez, J. M. G., Jörgensen, N., Törngren, M., Inam, R., Berezovskyi, A., Feng, L., Fersman, E., Ramli, M. R., & Tan, K. (2022). Edge computing for cyber-physical systems: A systematic mapping study emphasizing trustworthiness. *ACM Transactions on Cyber-Physical Systems (TCPS)*, *6*(3), 1-28.

66. Shukla, A., Katt, B., Nweke, L. O., Yeng, P. K., & Weldehawaryat, G. K. (2022). System security assurance: A systematic literature review. *Computer Science Review*, *45*, 100496.

67. Singh, R., & Gill, S. S. (2023). Edge AI: a survey. *Internet of Things and Cyber-Physical Systems*, *3*, 71-92.

68. Sisinni, E., Saifullah, A., Han, S., Jennehag, U., & Gidlund, M. (2018). Industrial internet of things: Challenges, opportunities, and directions. *IEEE transactions on industrial informatics*, *14*(11), 4724-4734.

69. Situnayake, D., & Plunkett, J. (2023). *AI at the Edge*. O'Reilly Media, Inc.

70. Stratigopoulos, H. G., Spyrou, T., & Raptis, S. (2023). Testing and reliability of spiking neural networks: A review of the state-of-the-art. In *2023 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)* (pp. 1-8). IEEE.

71. Sulaman, S. M., Oručević-Alagić, A., Borg, M., Wnuk, K., Höst, M., & de La Vara, J. L. (2014, August). Development of Safety-Critical Software Systems Using Open Source Software--A Systematic Map. In *2014 40th EUROMICRO Conference on Software Engineering and Advanced Applications* (pp. 17-24). IEEE.

72. Sun, X., Yu, F. R., & Zhang, P. (2021). A survey on cyber-security of connected and autonomous vehicles (CAVs). *IEEE Transactions on Intelligent Transportation Systems*, *23*(7), 6240-6259.

73. Tambon, F., Laberge, G., An, L., Nikanjam, A., Mindom, P. S. N., Pequignot, Y., Khomh, F., Antoniol, G., Merlo, E. & Laviolette, F. (2022). How to certify machine learning based safety-critical systems? A systematic literature review. *Automated Software Engineering*, *29*(2), 38.

74. Torens, C., Juenger, F., Schirmer, S., Schopferer, S., Maienschein, T. D., & Dauer, J. C. (2022). Machine learning verification and safety for unmanned aircraft-a literature study. In *AIAA Scitech 2022 Forum* (p. 1133).

75. Vyhmeister, E., & Castane, G. G. (2024). When Industry meets trustworthy AI: a systematic review of AI for Industry 5.0. *arXiv preprint arXiv:2403.03061*.

76. Wang, X., Han, Y., Leung, V. C., Niyato, D., Yan, X., & Chen, X. (2020). Edge AI: Convergence of edge computing and artificial intelligence. Springer Nature.

77. Wohlin, C., Runeson, P., Neto, P. A. D. M. S., Engström, E., do Carmo Machado, I., & De Almeida, E. S. (2013). On the reliability of mapping studies in software engineering. *Journal of Systems and Software*, *86*(10), 2594-2610.

78. Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering* (pp. 1-10).

79. Wohlin, C., Kalinowski, M., Felizardo, K. R., & Mendes, E. (2022). Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. *Information and software technology*, *147*, 106908.

80. Xiao, Y., Jia, Y., Liu, C., Cheng, X., Yu, J., & Lv, W. (2019). Edge computing security: State of the art and challenges. *Proceedings of the IEEE*, *107*(8), 1608-1631.

81. Zarrin, J., Wen Phang, H., Babu Saheer, L., & Zarrin, B. (2021). Blockchain for decentralization of internet: prospects, trends, and challenges. *Cluster Computing*, *24*(4), 2841-2866.

82. Zhang, J., & Li, J. (2020). Testing and verification of neural-network-based safety-critical control software: A systematic literature review. *Information and Software Technology*, *123*, 106296.

83. Zhang, J., Li, J., & Oehmen, J. (2023). Robustness evaluation for safety-critical systems utilizing artificial neural network classifiers in operation: a survey. Available at SSRN: *https://ssrn.com/abstract=4513915*.