Article:
Somers, Richard J. orcid.org/0000-0002-1101-9722, Douthwaite, James A., Wagg, David J. et al. (2 more authors) (2023) Digital-twin-based testing for cyber–physical systems: a systematic literature review. Information and Software Technology, 156. 107145. ISSN 0950-5849

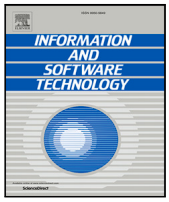
https://doi.org/10.1016/j.infsof.2022.107145

Information and Software Technology 156 (2023) 107145

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof



Digital-twin-based testing for cyber–physical systems: A systematic literature review

Richard J. Somers [a,*], James A. Douthwaite [b], David J. Wagg [c], Neil Walkinshaw [a], Robert M. Hierons [a]

[a] Department of Computer Science, The University of Sheffield, 211 Portobello, Sheffield, S1 4DP, United Kingdom
[b] Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield, S1 3JD, United Kingdom
[c] Dynamics Research Group, Department of Mechanical Engineering, University of Sheffield, Sheffield, United Kingdom

ARTICLE INFO

Keywords:
Digital twin
Cyber–physical system
Testing
Systematic literature review

ABSTRACT

Context: Cyber–physical systems present a challenge to testers, bringing complexity and scale to safety-critical and collaborative environments. Digital twins enhance these systems through data-driven and simulation based models coupled to physical systems to provide visualisation, predict future states and communication. Due to the coupling between digital and physical worlds, digital twins provide a new perspective into cyber–physical system testing.

Objective: The objectives of this study are to summarise the existing literature on digital-twin-based testing. We aim to uncover emerging areas of adoptions, the testing techniques used in these areas and identify future research areas.

Method: We conducted a systematic literature review which answered the following research questions: What cyber–physical systems are digital twins currently being used to test? How are test oracles defined for cyber–physical systems? What is the distribution of white-box, black-box and grey-box modelling techniques used for digital twins in the context of testing? How are test cases defined and how does this affect test inputs?

Results: We uncovered 26 relevant studies from 480 produced by searching with a curated search query. These studies showed an adoption of digital-twin-based testing following the introduction of digital twins in industry as well as the increasing accessibility of the technology. The oracles used in testing are the digital twin themselves and therefore rely on both system specification and data derivation. Cyber–physical systems are tested through passive testing techniques, as opposed to either active testing through test cases or predictive testing using digital twin prediction.

Conclusions: This review uncovers the existing areas in which digital twins are used to test cyber–physical systems as well as outlining future research areas in the field. We outline how the infancy of digital twins has affected their wide variety of definitions, emerging specialised testing and modelling techniques as well as the current lack of predictive ability.

1. Introduction

Cyber–physical systems present an ongoing challenge for engineers and testers. Increased autonomy, with increases in complexity and scale [1,2], provide a need for the safety and reliability procedures in cyber–physical systems to evolve in parallel. Testing techniques exist [2], but more comprehensive methods are required to ensure reliability in safety-critical situations. This review follows the definition of cyber–physical systems set out by Rajkumar et al. [1].

Digital twins are an emerging technology which combine time-evolving combinations of physics-based and data-driven models [3] with a coupling to a physical system to enhance its functionality [4]. Digital twins have been classified as "models", "shadows" and "twins" [5–7]. Under these classifications, models are assumed to provide visualisation of a physical system with no automatic data flow between the two. Digital shadows take real-time information about the physical system to present a virtual interface in a simulated environment. Digital twins close the feedback loop by combining digital models and shadows in order to supplement the physical system behaviour with greater insight into system information. This use of models attempts to support more autonomous execution of cyber–physical systems whilst enhancing their capabilities for safety and reliability [5]. Enhancements, such as introducing multi-physics simulation [8], visualisation of a physical

* Corresponding author.
E-mail address: rsomers1@sheffield.ac.uk (R.J. Somers).

https://doi.org/10.1016/j.infsof.2022.107145
Received 29 July 2022; Received in revised form 15 November 2022; Accepted 21 December 2022
Available online 26 December 2022

entity [5] and the ability for proactive state prediction for understanding the remaining life of a system [4], allows for more efficient use of a cyber–physical system throughout its life-cycle.

Digital twins can be used to provide a new perspective into testing cyber–physical systems. Integration with a system from design to decommission [9] can provide an adaptive and high fidelity model which acts as a test oracle for cyber–physical system behaviour. This can allow for testing through observing behaviour deviation from the digital twin to discover potentially erroneous behaviour as well as hardware degradation [10]. Digital-twin-based visualisation techniques enhance this further by providing detailed information about a system under test in real-time [5].

This paper presents a systematic literature review of existing techniques in which digital twins are used to test cyber–physical systems. We identify missing areas of research as well as speculate and propose the next steps required in this field. We used a manual systematic approach to reduce 480 initial studies to 26 studies that fit the defined review protocol.

This paper makes the following observations:

- Digital twin definitions are wide ranging and often domain specific causing confusion due to a lack of consensus. Across the many different areas of application, we explore multiple definitions and the confusion this lack of consensus has caused.
- Digital-twin-based testing is expanding as access to the technology becomes easier. Emerging platforms and frameworks [11] increase adoption and allow expansion into new areas.
- Digital twins in this review tend to favour passive testing techniques [12] which monitor the system during use. Moving towards active and predictive testing may provide more confidence in safety-critical systems.

The remainder of this review is laid out as follows: Sections 2, 3, 4 and 5 outline the background of the review including an overview of cyber–physical systems, digital twins and the other context required. Section 6 provides the methodology and protocol, as well as providing the research questions. Section 7 explores the results of the research questions and outlines any trends found. Section 8 discusses the results and trends to find areas of interest and areas which are missing. Finally, Section 9 concludes the review's findings and speculates about future uses in this area.

2. Cyber–physical systems: A testing challenge

This section presents an overview of cyber–physical systems, how they are designed and their current testing techniques. We use this section to highlight the complexity and challenges of testing cyber–physical systems.

2.1. Defining cyber–physical systems

Cyber–physical systems are software driven systems which "interact with the physical world, and must operate dependably, safely, securely, and efficiently and in real-time" [1]. These systems come in all ranges of autonomous ability which will be covered in this section. Cyber–physical systems are well established as a concept and have been widely explored by other systematic reviews [13,14].

To help our understanding of cyber–physical systems, we used the 5C taxonomy proposed by Bagheri et al. [15] for defining and designing a cyber–physical system at different levels of autonomous ability. This taxonomy is referred to during the remainder of the review to explore the different types of cyber–physical systems under test. A brief summary of the 5C taxonomy:

- **Connection** - Cyber–physical systems allow for a transfer of data between the physical environment and virtual software elements. Sensors allow for the virtual aspects of the system to obtain physical data and use it as part of the control process. This level is seen in the majority of cyber–physical systems.

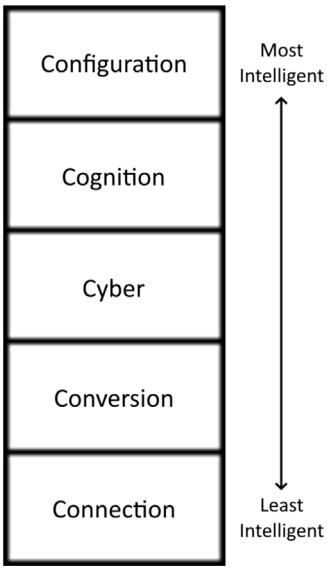


Fig. 1. 5C taxonomy levels of cyber–physical system intelligence.

- **Conversion** - The data gathered at the connection level can be converted into more useful, machine-understandable information for the system. Whether this is temperature data for health management or motor speed for later visualisation, interpretation of the physical data into something useful, such as an overall system health value, allows for a more informed awareness of the physical system.
- **Cyber** - A centralised hub for all information gathered by the cyber–physical system is known as the cyber layer. This non-physical layer allows for calculation and identification of erroneous behaviour in the system as well as comparison to past iterations and states of the system.
- **Cognition** - For data about a system to be used in an intelligent and informed way, it must be understandable and provide support to users of the system, whether they be humans or other systems. This level allows for visualised simulation of the system in a cyber context to better inform users of its behaviour.
- **Configuration** - This level of the cyber–physical system allows for self-adaptation using the data acquired and synthesised in the previous levels. This ability to change behaviour allows for resilience and optimisation to the current physical environment.

Fig. 1 presents the different levels of the 5C taxonomy compared to a cyber–physical system's intelligence. Systems start with the simplest layer, connection, and adopt further levels the more intelligent they are. Only very few cyber–physical systems achieve configuration. We return to this taxonomy in Section 3 when outlining how digital twins can be used to enhance cyber–physical systems.

2.2. Testing cyber–physical systems

The combination of both cyber and physical components and the dynamic environments in which they exist makes testing these systems incredibly difficult [16]. Four different testing objectives are identified by Zhou et al. [2]. We use these definitions to understand the types of testing found in this review. Each testing objective is outlined below:

- **Conformance testing** - Testing that a cyber–physical system conforms to the behaviour of some oracle. This can be implemented using a model and creating tests to ensure that the physical system conforms to an expected behaviour [17].

- **Robustness testing** - Investigating whether a cyber–physical system reacting to its stochastic environment will be classified as erroneous behaviour [18]. This is particularly important in cyber–physical systems as their physical environment will change over time, and this should not cause test failures.
- **Security testing** - Ensuring that a system cannot be affected by cyber-attacks which could either change the behaviour of the system or allow for system information to be intercepted. Due to the connection level of cyber–physical systems requiring networks, security testing is important to ensure they are safe and secure [19], especially when conducting safety critical tasks.
- **Fragility testing** - Ensuring that minor changes to inputs or the environment do not cause the system to completely stop working. Environmental uncertainties can cause minor changes in sensor readings due to being physical systems and this should not drastically affect system behaviour [2].

When exploring cyber–physical system testing, it is important to mention hardware-in-the-loop testing [20] as it is a precursor to digital-twin-based testing. Hardware-in-the-loop testing is used to test a specific physical component of a system while simulating the remainder of the system. Modelling the rest of the system allows for more specialised testing of a specific component while still testing it in the context of the entire system [20]. The difference between simulating this type of model of a system and digital twins is covered in Section 3.

3. Digital twins: Enhancing cyber–physical testing

Cyber–physical systems can be enhanced by providing sensor data to a software representation of themselves to allow for visualisation, computation and even prediction of the physical system's behaviour [4]. This representation is a system's digital twin. In this section we outline different definitions of digital twins, their classifications and how they are able to enhance existing cyber–physical systems.

3.1. Defining digital twins

How the physical system should be represented or modelled using a digital twin has a wide variety of definitions.

There is a lack of consensus on the definitions for digital twins. These definitions range from West et al. proposing that "integrated multiphysics, multiscale, probabilistic simulations" [8] should be used to others stating that how the model is developed does not matter, so long as it produces an output which is directly equivalent to its physical system [21].

Fuller et al. [6] examine six definitions and compare their ambiguity. They propose that an updated definition be developed to ensure digital twins are correctly identified and distinguished from other software models, as no single definition adequately distinguishes them. This is further covered in Section 6.2.5.

Ensuring that a precise definition is used is crucial in making sure our review is truly systematic and reproducible. This review follows the digital twin definition set out by Eyre et al. [4] wherein a digital twin is "a live digital coupling of the state of a physical asset or process to a virtual representation with a functional output". This definition is stated as Definition 1. For the remainder of this review, the physical system which is coupled to the digital twin will be referred to as the "physical asset" to follow this definition. The current lack of a singular definition provided an obstacle when finding relevant literature as it introduced many studies which did not fit our agreed definition.

Definition 1 (*Digital Twin*). A live digital coupling of the state of a physical asset or process with a virtual representation with a functional output.

3.2. Classifying digital twins

Classifying different types of digital twins is important for understanding their capabilities and the context in which they should be used. Due to digital twin definitions being so diffuse, as further explored in Section 8.1.1, there is no consensus on how to classify digital twins. This had lead to multiple different ways of doing so, with overlap in these taxonomies. This notion is extended to Section 3.3. To provide an overview of the different taxonomies used to classify digital twins, we explore two different classification methods in this section.

Eyre et al. [4] proposes three classifications for digital twins based on their most complex functional output. These classifications are used throughout the remainder of the review to examine the functionality of digital twins in testing. The classifications are outlined below:

- **Supervisory** - Supervisory digital twins accept data from the physical asset to provide information to a human observer, allowing them to act on this information. The functional output for this classification of digital twin is simply visual information.
- **Interactive** - Interactive digital twins close the feedback loop by using data gathered from the physical asset to make changes to the system in a time-evolutionary manner based on its current and historical states. Interactive digital twins have a functional output which can affect the physical asset.
- **Predictive** - Predictive digital twins predict the future states of systems by using information gathered in real-time as well as infer additional information for unseen contexts through physics based or data driven inference. The digital twin can use these predictions to change the physical asset's behaviour or issue preemptive warnings to human operators.

Douthwaite et al. [5] proposes a digital twinning framework with three different classifications for interacting with a physical asset. These are similar to other definitions found elsewhere in the literature [6,7]. We explore this taxonomy to provide a more informed definition of digital twins for the remainder of the review. Fig. 2 presents the data flow between the physical and digital systems for each of them.

- **Digital model** - Modelling provides a system with no automatic data exchange between the simulation and the physical asset so the simulation can run independently. This classification is outside our definition of a digital twin but importantly highlights how model-based approaches are used for cyber–physical systems.
- **Digital shadow** - Digital shadowing allows a simulation to mimic a physical asset where data is passed only from the physical asset to the digital twin in real-time. This can provide a "visual representation" [4] of the system. This classification provides behaviour equivalent to the supervisory classification above.
- **Digital twin** - The final classification is that of the digital twin itself where both the physical asset and digital twin exchange data to allow for real-time analysis, interaction and adaptive behaviour. This classification closes the feedback loop allowing for interactive digital twins with the ability to provide predictive capabilities where necessary.

Both classification systems outlined above have overlap but present different perspectives on the digital twin. The classification set out by Eyre et al. [4] focuses primarily on the functional output of the digital twin whereas Douthwaite et al. [5] explores the way in which information is shared within the system.

3.3. Enhancing cyber–physical systems

The introduction of digital twins into cyber–physical systems is closely related to the concept of Industry 4.0 [9]. This includes the use of digital twins to enhance the performance and efficiency of

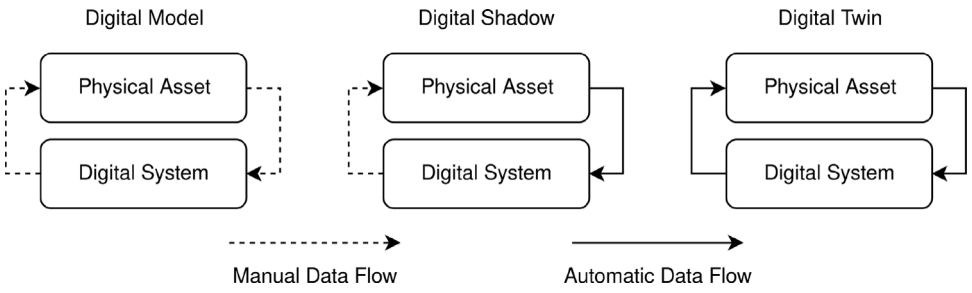


Fig. 2. Digital model, digital shadow and digital twin.

cyber–physical systems by introducing multi-physics simulation [8], visualisation of a physical entity [5] and the ability for proactive state prediction for understanding the remaining life of a system [4]. Adoption of digital twins enables more autonomous cyber–physical systems with self-adaptive configurations, as seen in autonomous vehicles [22], to ensure safety and correct behaviour. In addition, these capabilities support the adoption of more complex systems of systems. Digital twins allow the modelling of more intricate interactions, such as those between individual cyber–physical systems in smart cities [23].

The 5C taxonomy for cyber–physical systems [15], summarised in Section 2.1, can be enhanced at the higher levels by the use of digital twins. As mentioned in Section 2.1, we refer to these definitions throughout the remainder of the review. These levels are outlined in a digital twin context below:

- **Cyber** - This layer can act as the system's digital twin as it has access to all information passing though the system as well as all its past states. This information is used to generate a model of the system which runs in parallel with the physical asset.
- **Cognition** - Digital twins enhance this level by providing granular real-time visualisations of the system with comprehensive information generated by the rest of the cognition level.
- **Configuration** - The full potential of digital twins is present at this level. The closed feedback loop can adapt a system's configuration to produce safer and more advised behaviour with further improvement from using predictive behaviours.

Since the cyber level above only provides a vague interpretation of how a cyber–physical system can contain a digital twin, we explore another, more precise taxonomy. A classification of digital twins is defined by Bonney et al. [24] which breaks these systems down into 3 layers. This taxonomy focuses on the construction of the digital twin, as opposed to the capabilities outlined by the 5Cs [15]. This taxonomy is not referred to in the research questions but is included for completeness. A brief summary of this taxonomy:

- **Internet-of-things layer** - This layer contains the communication between the physical asset and the digital space. Sensors, control hardware and actuators are all contained in this layer.
- **Interface layer** - To provide enhancements to a cyber–physical system, the interface layer coordinates behaviour for required tasks and workflows based on information gathered by the internet-of-things layer. This layer also allows communication between network services, seen in the cloud computing layer, and the internet-of-things layer.
- **Cloud computing layer** - Network services such as cloud data storage, high performance computing and other remote aspects relating to the digital twin make up the cloud computing layer. They provide enhancements to the physical and digital systems through additional network-based services.

3.4. Digital twin oracles

As mentioned in Section 2.2, hardware-in-the-loop testing is different to testing techniques which use digital twins. Digital twins allow for an entire system to be modelled "live" [4] as a coupled entity to allow for system verification throughout the system's lifetime. Hardware-in-the-loop instead models only certain parts of the system during development [20]. This difference allows for a more rigorous and ongoing testing procedure. The user can achieve testing of the complete system as well as connections between digital twins.

Digital twins simulate indistinguishable behaviour to that of the physical asset. Grieves [25] suggests the "Grieves Performance Test" where a human is asked to decipher the difference between the output of a physical asset and that of a simulation. A similar test is presented by Worden et al. [26] where a Turing Mirror is used to provide test cases to either a physical asset or a simulation. These tests are intended to provide insight into the success of simulating physical assets based on their accuracy. Grieves goes as far to say that simulations can be more useful in testing due to the physical limitations of testing environments. These tests show that comparison to a representation of a system is not a new concept, but has only recently been introduced to cyber–physical systems [9].

4. Definitions

This review uses general testing terminology and different contexts throughout the following sections. This section defines those terms to help provide a full understanding of their use throughout the remainder of this review.

Definition 2 (*Modelling Context*)**.** A model is a simplified representation of another system to typically provide analysis and verification external to the original system [27]. The context of modelling in digital twins is further explored below.

Definition 3 (*Testing Context*)**.** System testing is used to discover problems with a system and assess the quality of said system. Testing is required to ensure the human element of programming and system design does not negatively affect the systems performance [28]. The context of testing in digital twins is further explored below.

Definition 4 (*Test*)**.** The act of executing test cases with the goals of finding failures in execution as well as determining correct functionality [28].

Definition 5 (*Test Case*)**.** An expression of a particular piece of program behaviour with a set of inputs and expected outcomes [28].

Definition 6 (*Test Oracle*)**.** A process of defining whether a test case being executed provides behaviour that is correct or not [29].

Definition 7 (*Failure*)**.** A state in which a system does not behave in its intended manner. This is typically observed by other system or a human user [30].

Definition 8 (*Digital-twin-based Testing*)**.** Testing which is performed on a system by the digital twin of the physical asset under test.

4.1. Contexts

The definition of some terminology is dependent on its context. Since this paper deals with both modelling and testing contexts, the terms described below are defined in both contexts.

This paper explores models of systems through the use of digital twins. Wagg et al. [3] presents classifications of digital twins as an extension of verification and validation techniques that make use of the model classifications: "white-box models", "black-box models" and "grey-box models". The following definitions explore these terms from an engineering modelling context:

Definition 9 (*White-box Modelling*)**.** A model built with physics based reasoning. White-box models provide visibility of information within the model. They are models which are easily understood by experts in associated fields and provide explanations as to how an outcome has been derived [31].

Definition 10 (*Black-box Modelling*)**.** A model derived purely from data with no physics reasoning. Black-box models provide visibility of only the input and outputs and no visibility of information inside the model. They are models and methodologies which are difficult to understand, even for experts in the field, and provide no explanation as to how model outcomes are derived [31].

Definition 11 (*Grey-box Modelling*)**.** Grey-box models provide visibility of inputs and outputs as well as partial visibility of information within the model [32]. They provide a combination of white-box and black-box to provide some understanding of a model as well as a partial explanation as to how outcomes are derived.

As well as a modelling context, "white-box", "black-box" and "grey-box" are terms widely used in software testing. As this paper explores testing techniques, we also define these terms in a software testing context:

Definition 12 (*White-box Testing*)**.** Testing through a detailed investigation into the internal workings of the system under test. Techniques require access to the source code for this to be applicable [33].

Definition 13 (*Black-box Testing*)**.** Testing which does not investigate the internal structure of the system under test but instead focuses on inputs and outputs of the system. The source code of the system has little to no relevance to this type of testing so it does not need to be accessible [33].

Definition 14 (*Grey-box Testing*)**.** A combination of both white-box testing and black-box testing techniques. This is used when there is some access to the internal structure of the system under test, but it is also supplemented by inputs and outputs [33].

These terms are used in both modelling and testing contexts throughout the remainder of this review. The context in which they are used is stated alongside them.

5. Related systematic reviews

There does not appear to be a systematic literature review specifically for testing techniques that use digital twins to test cyber–physical systems. There are, however, literature reviews for related topics.

Zhou et al. [2] conducted a literature review on the different testing techniques for cyber–physical systems which focused on the testing objectives of different testing techniques, how cyber–physical system testbeds are generated, and future research into cyber–physical system testbeds. Digital twins are mentioned briefly during the section on cloud-based testing but are not further explored whereas real-time monitoring and verification are covered through the use of hardware-in-the-loop and model-in-the-loop testing techniques.

Systematic literature reviews focused on digital twins are present in the literature focusing on the characterisation of digital twins [34] as well as exploring digital twin applications [35]. A systematic literature review on the applications of digital twins was conducted by Semeraro et al. [35]. This review explains how current digital twins are used in industry, their function within those industries, and how their architecture is defined. Using digital twins to test cyber–physical systems in real-time is again briefly mentioned during this review as part of the digital twin life cycle but not further explored. Our review covers a function of digital twins in industry with their application into testing their cyber–physical twin systems which fits into the research challenge of the applications of digital twins.

The background of digital twins and their role within cyber–physical systems as well as the current testing techniques available to cyber–physical systems are explored in these related studies [2,34,35]. Our review covers the combination of both areas in which techniques using the digital twin are used to test the cyber–physical system.

6. Research methodology

In this section, we start by defining the research questions. We then outline the methodology used to create a search protocol and retrieve studies to answer our questions in Section 7.

6.1. Research questions

Four research questions are outlined in this section. The questions are designed to produce a broad overview of existing testing as well as provide an understanding of the current capabilities of digital-twin-based testing.

6.1.1. RQ1: What cyber–physical systems are digital twins currently being used to test?

As mentioned in Section 1, cyber–physical systems are being enhanced by digital twins in the context of Industry 4.0 [9]. This questions aims to observe the introduction of digital twins and how this introduction is affecting different technologies and engineering industries in a testing context. In outlining this, primary areas, emerging areas and gaps in digital twin implementation can be found. This will also provide insight into whether different areas have different testing requirements and the possible reasoning behind them.

6.1.2. RQ2: How are test oracles defined for cyber–physical systems?

A test oracle defines the desired and undesired behaviours of a system. Cyber–physical system behaviour can be tested against an oracle to investigate whether it is behaving correctly. This question uncovers how test oracles for digital-twin-based cyber–physical system testing are defined. As some digital twins allow for adaptive systems whose behaviour may change depending on the environment, it is important to understand how test oracles in this field account for this potential unpredictability.

6.1.3. RQ3: What is the distribution of white-box, black-box and grey-box modelling techniques used for digital twins in the context of testing?

This question is used to determine whether one modelling technique is more commonly used when testing cyber–physical systems with digital twins, or if other factors affect this decision. As this question focuses on how the digital twin is designed, we approach white-box, black-box and grey-box from a modelling context to provide insight into the information visibility throughout the digital twin as well as whether the model is derived from physics or is data driven [3,32].

Table 1
Digital libraries used to retrieve studies.

Digital library	URL
ACM	https://dl.acm.org/
IEEE	https://ieeexplore.ieee.org/
ScienceDirect	https://www.sciencedirect.com/
Scopus	https://www.scopus.com/

6.1.4. RQ4: How are test cases defined and how does this affect test inputs?

Test cases capture different scenarios in which intended and unintended behaviour is explored. These scenarios ensure that correct behaviour is exhibited by the system. This question explores test cases to investigate how different testing techniques, such as specific test scenarios and continuous behaviour verification, are used in digital-twin-based testing of cyber–physical systems. Different approaches to selecting test inputs are also explored in this question to understand how different test case approaches affect this selection. This question will outline how a digital twin and its cyber–physical system interact in a test based environment and provide information about the data transfer required for testing.

6.2. Review protocol

A review protocol is developed in accordance with the steps set out by Kitchenham et al. [36] to produce a systematic literature review. We use these guidelines to ensure the review is both unbiased and reproducible. Other techniques were considered such as a systematic mapping review, as defined by Peterson et al. [37], as well as a multi-vocal literature review, as defined by Garousi et al. [38]. We decided to produce a systematic literature review as a lack of exploration into the subject area, defined in Section 5, would have caused difficulty in developing the mapping review as well as a lack of relevant grey literature, explored further in Section 6.2.5, did not allow for a multi-vocal literature review to be produced.

6.2.1. Digital libraries

Table 1 shows the digital libraries used to retrieve studies during this review. We chose these four digital libraries because of their reputation, ability to process advanced search queries as well as availability of studies for the given subject. To ensure that the digital libraries chosen were inclusive enough, we produced a test set of studies. This test set included studies we were sure fit the criteria of the review and should be included in the initial searches. We tested each study against the digital libraries to ensure each was present in at least one. This guaranteed the availability of relevant studies.

6.2.2. Search strategy

To successfully search the digital libraries outlined in Section 6.2.1, a search query was outlined using the PICO method, as suggested by Kitchenham et al. [36]. The PICO method outlines the Population, Intervention, Comparison and Outcomes of the research question to define a search query.

- **Population** - The type of technologies or users which the review is obtaining data about. In this case, the population is the cyber–physical systems under test.
- **Intervention** - The methodologies or tools used to address the specific issue. The intervention for this review is digital twins as they are being investigated as to how they are being used to test cyber–physical systems.
- **Comparison** - The methodologies or tools which the above intervention is being compared to. In this case, there is no comparison as we are only investigating digital-twin-based testing. A comparison, such as model based testing, could have been used to see the differences between the two testing techniques.

```
(Cyber-Physical OR Physical OR Hardware) AND
(System OR Device) AND
("Digital Twin") AND
(Test OR Validation OR Verification)
```

Fig. 3. Final search query used in digital libraries to find relevant literature.

Table 2
Inclusion, exclusion and quality criteria.

ID	Criteria
IC1	A least one testing technique is described
IC2	The system under test must be a cyber–physical system
IC3	Testing is performed using a digital twin
EC1	The digital twin described does not use a live data coupling
EC2	The study describes future use of a digital twin
EC3	Non-english study
EC4	Not published in a journal or conference proceedings
QC1	Are the research questions of the examined study answered?
QC2	Is the study reproducible?

- **Outcomes** - The overall outcome of the intervention. As we investigated how digital twins are used in testing, the testing and validation technique used on the system is used as the outcome.

Using the above definitions, a search string was developed by combining sections with "AND" statements as well as "OR" statements between each word within the section. The population of this search string has been split across two sections to allow for more variety in capturing how a study may define a cyber–physical system. The final search string is shown in Fig. 3.

This search string was tested to ensure it captured the relevant studies by using the test set of studies, outlined in Section 6.2.1, which we knew it should encompass. The search string was refined over several iterations to produce the one presented in Fig. 3.

6.2.3. Search automation

As part of the review protocol, Tsafnat et al. [39] proposes possible automation techniques to improve the efficiency of performing a review. This section outlines the search automation attempts made throughout the review.

As mentioned in Section 6.2.2, creating the correct search string took multiple iterations with different search queries. We developed a puppeteer client to crawl the digital libraries and perform the search query to return the relevant studies. The test set, outlined in Section 6.2.1, was then compared to the results to ensure the search query produced relevant literature and was not too vague. Our final search string was found by attempting different iterations of the search string using this method.

After all the studies had been selected, a Python program was developed to extract metadata from their bibtex files. This process removed duplicates as well as created a csv file of all metadata for keeping track of studies. This csv was used to keep track of all acceptance or rejection criteria for each study during each section of the review process specified in Fig. 4. This csv was converted into an Excel file to allow for more comprehensive formatting and formulas and has been made available.[1] Minor edits had to be made to this data for it to comply with copyright, so the abstracts and keywords of the studies have been removed.

6.2.4. Selection criteria

We devised specific criteria for both accepting and rejecting studies to ensure only relevant literature was accepted into the review. For

[1] https://doi.org/10.15131/shef.data.19383521.

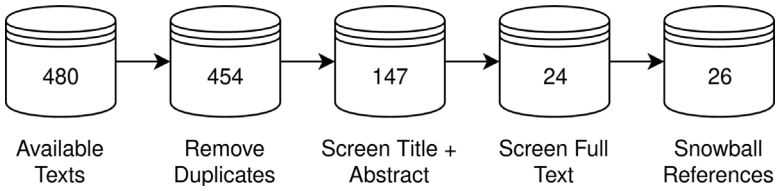


Fig. 4. Number of studies throughout the search protocol.

a study to be accepted, it must fit all the inclusion criteria, include no exclusion criteria and pass all quality criteria. This system allowed for a strict yet systematic approach to finding relevant materials while helping to reduce bias. The criteria used are shown in Table 2.

Digital twins having a live data coupling (EC1) was included to ensure that the digital twins described in the study were digital twins as described by Eyre et al. [4], which is the definition this review follows as explained in Section 3 and presented in Definition 1. As covered in Section 6.2.5, the disparity in agreement of the definition of a digital twin meant that it was important to include this exclusion criteria so that model based testing techniques [40] were not included in the review.

Ensuring studies contained cyber–physical systems (IC2) was challenging and required a strict definition to ensure the validity of the review. Systems which contained a software component which interacted with a physical component were accepted as part of this study. This allowed for systems such as normal pumps to be excluded from this review, but pumps which had software component to manage the system could be included.

The quality criteria selected ensured that the studies accepted into this review answered the questions they set out to answer (QC1) and appeared to be reproducible (QC2). We deemed a study reproducible if it provided sufficient data, a detailed methodology and measurable results which would allow for the study to be reproduced. These criteria allowed for studies, which appeared relevant but did not provide enough information to answer the research questions, to be rejected.

6.2.5. Grey literature

With the topic of digital twins still in its infancy, additional grey literature was explored to provide a greater perspective into such an emerging technology. Following the steps outlined by Garousi et al. [38], a multi-vocal literature review was attempted. As described in this study, grey literature can be found in multiple different ways. Search engines and exploring digital twin specific conferences revealed nothing useful so instead industrial experts were consulted for their opinion on finding such literature.

A few examples of commercial digital-twin-based projects emerged from this discussion but this only uncovered literature surrounding tools instead of commercial testing implementations. One such example, AWS Twin Maker [11], allowed for user specified digital-twin-based testing within the tool. We did not include grey literature surrounding these tools as part of the study as they did not answer our research questions, but the interest garnered by industrial experts around them showed their impact on the evolving subject of digital twins. Although not relevant to this review, we believe that further work should be performed into exploring the support provided by these tools for testing cyber–physical systems.

The British standard for digital twins [41] was revealed as a recent piece of grey literature of interest mentioned by most experts. However, it appeared to be contentious due to being too restrictive. Due to digital twins still being an emerging technology, the consensus reached was that grey literature would more likely mislead the review rather than provide useful input due to a lack of a concrete definition of digital twins. Because of this, grey literature is not included as part of this review.

Table 3
Study search results and availability.

Digital Library	URL	Results	Available
ACM	https://dl.acm.org/	209	204
IEEE	https://ieeexplore.ieee.org/	125	124
ScienceDirect	https://www.sciencedirect.com/	90	85
Scopus	https://www.scopus.com/	251	67
Total	–	675	480

7. Results

This section explores the execution of the search protocol as well as the results to each of the research questions. A separate bibliography is used during this section to refer to the studies found in this review.

7.1. Query execution

After finalising the search string, we executed the query against the digital libraries outlined in Section 6.2.1. Table 3 presents the results for each digital library after the initial search and the number of studies available in full text. Some digital libraries showed discrepancies between their search results and their study availability. We found this was due to digital libraries showing results for papers only available on other digital libraries, inflating their search results with duplicates. All studies were ultimately available as they were downloadable through their original digital libraries.

After obtaining the texts for each study, they were stored using Mendeley Reference Manager. This provided a centralised place where we could collaboratively access and annotate all the studies.

As is shown in Fig. 4, we identified 24 studies from the initial search before snowballing. We manually reduced 480 studies down to 24 by comparing them to one another through further duplicate screening, reading each abstract and finally reading the full text with regards to the inclusion and exclusion criteria. It took weeks to reduce these studies down to only relevant papers, due to all the false positives, and was the most time consuming part of the study.

Two studies were included during snowballing as they were missed by the initial query. This was due to specific reference to technologies which would have caused an overly specific and long search query had they been included. This problem was exacerbated by the maximum number of terms allowed when searching digital libraries as they did not allow for enough terms to provide an exhaustive search.

We updated the csv produced in Section 6.2.3 to include at which step in the query execution each paper was accepted or rejected. This process allowed us to keep a record of why each paper was accepted to further increase the reproducible of the search protocol.

7.2. RQ1: What cyber–physical systems are digital twins currently being used to test?

This section explores the different types of cyber–physical system under test. It was difficult to produce a grouping for all of the different areas of use of digital twins in testing due to the vastly different sectors found in our results. This required a taxonomy which provided easy to understand groupings with relevance to the systems found.

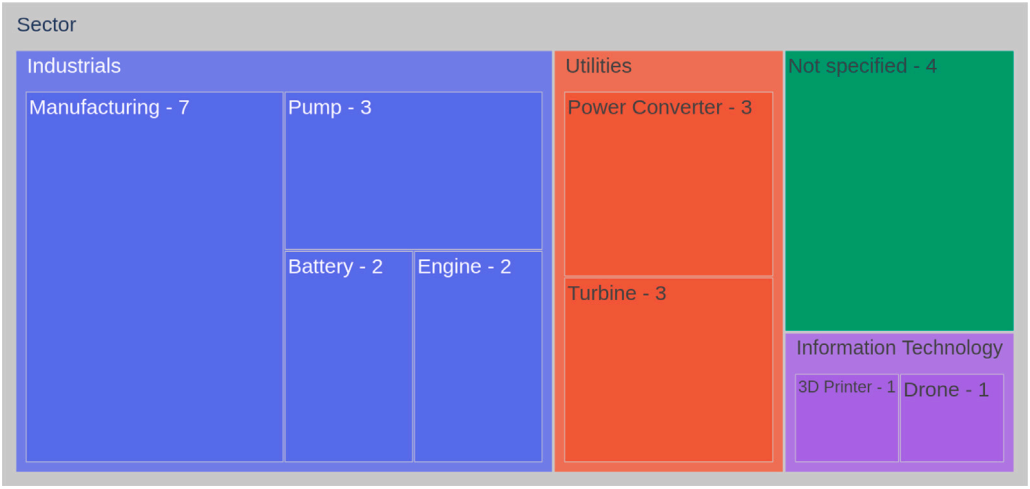


Fig. 5. The distribution of cyber–physical system types present across the studies in accordance to GICS sectors.

Table 4
Type of cyber–physical system present in each study.

Cyber-physical system	Studies
Industrials	
Manufacturing	[52–58]
Pump	[59–61]
Engine	[62,63]
Battery	[64,65]
Utilities	
Power converter	[66–68]
Turbine	[69–71]
Information technology	
Drone	[72]
3D printer	[73]
Not specified	[74–77]

The global industry classification standard (GICS) [42] was used to split different types of cyber–physical system into different sectors and explore how this may affect the testing techniques used. As is shown in Table 4, three different sectors were outlined with a fourth section for an unspecified cyber–physical system. The distribution of these systems within the sectors is shown in Fig. 5. This taxonomy was suitable due to the wide variety of cyber–physical systems it covered and the different levels of granularity that could be applied.

Some of the applications below, such as pumps and batteries, can be described as physical systems as opposed to cyber–physical systems. To ensure we conformed to IC2 from Section 6.2.4, our studies were restricted to applications which contained both software and physical components [1].

7.2.1. Industrials

The industrials sector includes systems in this review from the machinery, aerospace and defence as well as the construction and engineering industries [42]. This sector makes up the largest portion of studies in this review so represents the main sector in which digital twins are being used to test cyber–physical systems.

Manufacturing Manufacturing based systems are those which work, typically in factories, to create products. Cyber–physical systems see a large adoption into this area due to their ability to physically manipulate objects with the added connectivity and communication of their cyber components.

We found that manufacturing based applications of digital-twin-based testing were the most common in this review. They tend to be complex, modular systems with each section being required to work together and not interfere with one another [43]. Industry 4.0 adoption has pushed for the introduction of digital twins in manufacturing settings to help provide safety and system optimisation [5,9].

Some studies found in this review contained digital twins of entire manufacturing plants [52], whereas others described digital twins for each specific device within a manufacturing plant to create a framework of modular digital twins [53]. The way in which the cyber–physical system in manufacturing is tested appears to be the same across most studies; the physical asset is compared to its digital twin and the deviation in behaviour of the two is measured. This is done either as a classifier [54,55,52] or more specialised techniques, such as simulating events on the digital twin first to see if failures would occur [58] as well as ensuring no network tampering occurs through deviation from a digital twin state machine [53]. Although these approaches have differences, the main application of real-time monitoring of both the physical asset as well as its digital twin to ensure they do not go out of sync is always present. This is achieved by the digital twin defining the correct behaviour that the physical asset should be following.

Ensuring physical machinery are rigorously tested in manufacturing is essential for safe procedures in human machine collaborative spaces. Since manufacturing machinery is often large and potentially dangerous when acting incorrectly, it is unsurprising that this is the largest area of adoption found within this review.

Battery Battery based cyber–physical systems are those which allow for the storage and release of electricity with the software enhancement of control and monitoring. We found that this area contained studies from very different application settings with one belonging to the general electrical equipment industry [65] and the other belonging to the aerospace and defence industry [64].

Battery cyber–physical systems use a different type of testing compared to that seen in manufacturing. The approaches described in the literature [64,65] focus on degradation monitoring to better understand the life-cycle of a battery as opposed to its general failures. This digital twin testing technique allows for more information about how a battery will degrade over time to inform when they may need to be replaced or maintained. Battery based cyber–physical systems provide an example of how the predictive capabilities of a digital twin in combination with the cognition and configuration levels of cyber–physical systems can be used.

Engine Similar to both manufacturing and battery systems, engine cyber–physical systems use a mixture of real-time sensor monitoring to determine failures [62] as well as using classifiers to determine degradation over time [63]. Because of this, there appears to be less

specialisation into the testing of engines with digital twin compared to the previous pump examples even though they both make up the same proportion of adoption.

Both types of testing found in this area make sense for engines due to their complex nature as a critical cyber–physical system. They require safety through failure detection, as well as contain large number of moving parts which causing degradation of components. Digital twins are used to provide testing for both of these characteristics, but do not provide a specialised testing technique for engines.

Pump Pump systems make up a considerable part of the industrials sector studies found with this review with very different applications. Yoginath et al. [59] explores a canal lock system run by pumps, Lu et al. [60] shows a heating, ventilation and air conditioning system with a pump as the cyber–physical system under test, and Short et al. [61] proposes a pump as an example of a cyber–physical edge device to be tested. Despite the different applications of the pumps, the digital-twin-based testing method is the same in each study by testing for deviations of the physical asset from that of its digital twin, as seen in manufacturing.

As pumps have moving parts, it is interesting to not see any degradation testing in the studies. Since digital twin testing is still in its infancy, we believe that more rigorous and predictive testing will be seen in pumps through the use of digital twins in the future.

7.2.2. Utilities

The utilities sector provides access to energy, gas and water [42]. This review includes systems from this sector which are industrial power converters and wind turbines, making up the second largest sector found.

Power converter Power converters are cyber–physical systems which convert electricity between AC and DC whilst providing an uninterrupted flow of electricity [44]. Testing power converters is important to ensuring that no failures occur due to incorrect voltage regulation to protect both the users, as well as any machinery connected.

Digital twins can be seen in the studies to provide this kind of fault tolerance by following similar testing techniques to manufacturing. Sensor deviation is used to see how voltage, current and other sensor data, such as temperature, can be used to detect failures as the physical asset produces different sensor data to that modelled by its digital twin [66–68]. These systems incorporate interactive digital twins with no levels of predictive behaviour.

Turbine Wind turbines follow the testing techniques explored in engines. We did not find this surprising due to the similarity of moving components causing wear as well as their critical nature requiring failure detection.

A mixture of classifiers to determine degradation of bearings over time [69] as well as sensor deviation of the physical asset from its digital twin [70,71] techniques are found in these studies. The inclusion of degradation testing emphasises how digital twins enhance testing when moving parts are present by allowing for more comprehensive, real-time modelling.

7.2.3. Information technology

The information technology sector contains systems in this review which are technology hardware and equipment [42]. This sector makes up the smallest section of digital twin tested cyber–physical systems found in this review. Similar to manufacturing, pumps and power converters, both examples found in the primary studies of this review use a form of real system deviation from the digital twin to define erroneous behaviour.

Drone Use of digital twins to test a drone system is shown in Grigoropoulos et al. [72] by comparing the state of the physical asset with that of the simulated digital twin. Discrepancies are identified from this state difference as system failures to be corrected.

Table 5
Oracle types found within each study.

Oracle type	Studies
Specified	[72,62,53,77,73,76,65,66,57,68,61,58,67]
Derived	[70,54,75,55,69,63,60,64,52,74,59,71]
Implicit	
Human	[56]

3D printer Henson et al. [73] performs testing in a similar way to that seen in drones. This was performed by comparing a generated model based on the input file with cameras around the 3d printer to ensure a correct printing procedure.

The use of digital twins in drones and 3d printing appears to be an emerging area of digital twin application. This is due to the limited literature surrounding digital-twin-based testing in these areas. Digital twins provide the configuration level of a cyber–physical system to drones, as well as providing comprehensive conformance testing to 3d printing.

7.2.4. Not specified

Non specialised systems described in the literature express testing cyber–physical systems without a specific system in mind. These studies proposed general purpose solutions to provide testing using digital twins without the limitations of a specific system.

Similar to manufacturing specific system testing, deviation between the physical asset and its digital twin [74,75,77] is a common application in non system specific testing. Some more specialised applications of digital twin testing are also defined. Gao et al. [75] does not define the test classifier as part of the digital twin but instead describe a framework which can be used to test any system and its digital twin. Li et al. [76] provides a novel approach by comparing the timing of sensor data as opposed to the data itself to find failures. As conformance testing appears to be the standard across the other sectors, it is not surprising that non specialised techniques follow this.

7.2.5. Summary

This section has explored the many different classifications of cyber–physical systems currently being tested by digital twins. A focus on conformance testing, as outlined in Section 2.2, was found in these studies with only a few examples of other types. As a result of this conformance testing, it can be seen that these cyber–physical systems exhibit behaviours related to both the cognition and configuration levels, outlined in Sections 2.1 and 3.3.

7.3. RQ2: How are test oracles defined for cyber–physical systems?

For a test to either pass or fail, a test oracle must describe the correct or incorrect behaviour of the system under test. Barr et al. [29] define four different types of test oracle: “specified”, “derived”, “implicit” and “human”. As is shown in Table 5, three out of the four types of oracle are found within the studies in this review.

7.3.1. Specified

Specified oracles are those which are built based on the specification of the system to ensure that it conforms to its pre-specified correct behaviour. In the case of digital twins, a twin built within the specification of its cyber–physical system can act as the oracle during testing. Specified oracles make up the largest group found within the studies in this review.

Grigoropoulos et al. [72] uses “off the shelf” drone simulation technology to specify how the digital twin of a real drone should act. This acts as a specified oracle as deviation in the real system from this simulation software can be seen as erroneous behaviour. Other approaches such as black-box specification generated digital twins [62,77] and state machine based digital twins [53] shy away from fully

simulating the system, but provide the same insight into if a physical asset is deviating from its specification.

Defining the specification can be done in multiple ways; as mentioned above, modelling using specialised software provides a simulation that the physical asset should follow. Xia et al. [58] use computer aided design software to justify what a physical asset is able to do and how the components within that system will react to movement and changes by other components. If the components in the real system do not act and react in the same way they do while being simulated, then a failure has occurred.

Specified oracles do not always require complex simulation software to be achieved. Henson et al. [73] uses the input file to be printed on a 3d printer as an oracle for the printing process. This approach uses multiple cameras ensuring that each layer of the print is within a threshold of the original source file.

7.3.2. Derived

Derived oracles are defined by existing behaviours and correctly functioning historical data of the system. Digital twins can be built using data of a physical asset that is working correctly to then mimic the system and find deviations from a physical asset and the original correctly functioning data set.

Further explored in Section 7.4.2, neural networks are used as digital twins to mimic correct behaviour of a system. These neural networks require training data to be created and this is either provided by purely historical system data [74,54,64,69] or a combination of historical data and ensuring it conforms to the system specification [52,70,55]. Including a combination of derived oracle and non-derived specification allows for more information about a failure to be obtained, providing specific failure classification information to a tester [70].

Most studies in this review have the digital twin as the oracle during testing but this is not always the case. Separate failure classifiers [75] show how historical data from a failure free system as well as that of a working digital twin can be used to train a failure classification oracle separate from the digital twin. An oracle defined like this both provides test failures as well as the classification of that failure which would aid in its investigation. This approach is further explored in Sections 7.4.1 and 7.4.3.

7.3.3. Implicit

Implicit oracles are those which find when something has objectively gone wrong. Creating these oracles can be very difficult as behaviour defined as simply incorrect for one system may be correct for another. For example, a program crashing could be seen as an implicit error for most programs, although this can be intended behaviour in a program is designed to crash [29]. No systems with implicit oracles were found in this review.

7.3.4. Human

Human defined oracles are those in which a human ultimately decides whether a behaviour is right or wrong. This approach is used when the creation of an oracle cannot be automated and a human opinion is required.

A single example of a human based oracle was found during this review. Kang et al. [56] uses a technique in which the oracle used to define correct behaviour by involving a human operator at each state of the digital twin and physical asset hierarchy. This ensures multiple levels of verification of the systems actions in a proactive manner to find failure prone behaviour. Having a human element provides additional cost and uncertainty as it can require further expertise in the system to understand which behaviours will lead to errors.

7.3.5. Summary

Digital twin's act as the oracles in the studies of this review. The majority of oracle types found were defined as either specified or derived oracles. The small number of human oracles suggests a shift away from human interaction in the testing process for a more automated approach.

Table 6
White-box, black-box and grey-box modelling across the studies.

Classification of model	Studies
White-box	
Simulation	[61,72,56,73,57,58]
Deterministic mathematical	[68,67,65]
Stochastic mathematical	[66]
State machine	[53]
Black-box	
Neural network	[69,71,74,52,59,64,76,77,60]
Grey-box	[70,75,62,54,55,63]

7.4. RQ3: What is the distribution of white-box, black-box and grey-box modelling techniques used for digital twins in the context of testing?

This section explores whether the technique used to model cyber–physical systems are white-box, black-box or grey-box as well as how test failure classification affects this. This provides insight into the data visibility throughout the model and how a user may be able to infer information throughout. Table 6 presents the studies and the techniques carried out within them and Fig. 6 further explores this distribution.

7.4.1. White-box

White-box models are models which provide an understandable representation of information transfer throughout the model which can be interpreted by experts in the field [31]. In the context of digital twins, white-box models have been described as a model "where the equations of motion have been derived from the underlying physics of the problem and the model parameters have direct physical meanings" [45]. As found in Section 7.2, testing is dependent on the digital twin and therefore the way in which the digital twin is constructed, as well as any classifiers involved, should be understandable for a white-box modelling classification to be applicable.

Simulation We define simulation based models as those which use equation based physics in combination with a visualisation to set them apart from mathematical based models. Simulation based digital twins provide a detailed implementation of the inner workings of the system under test. These are normally produced using specialised software to produce detailed simulations as close to reality as possible. These models act as white-box through their use of underlying physics and information visibility throughout the model [3,45]. Simulation based digital twins also tends to provide visualisation of the system in real-time to allow for the cognition level of cyber–physical systems to be achieved.

Short et al. [61] uses the design and analysis package MATLAB Simulink to provide an in-depth simulation of their system under test. Insight into the simulated system during a test can be gained and visualised to aid in finding failures due to a high level of information visibility. Grigoropoulos et al. [72] provides a virtual drone simulated alongside a physical drone. The software used [46] presents a physics based simulation and clarity into information transfer throughout the model. Historical logs are generated throughout the virtual and physical flights which can then be compared and used to aid in understanding test failures.

Deterministic mathematical Deterministic mathematical based digital twins simulate the system under test in the form of physics based equations. This type of digital twin typically does not provide the visualisation that simulation-based digital twins do but provides visibility of the information passing through the model [45]. This technique allows more concentrated computation on the system modelling as computation is not required for its representation to the user.

This approach appears in the studies when electronic components are modelled using existing component specific equations [68,67,65]. It is particularly useful in these studies as only current, voltage and

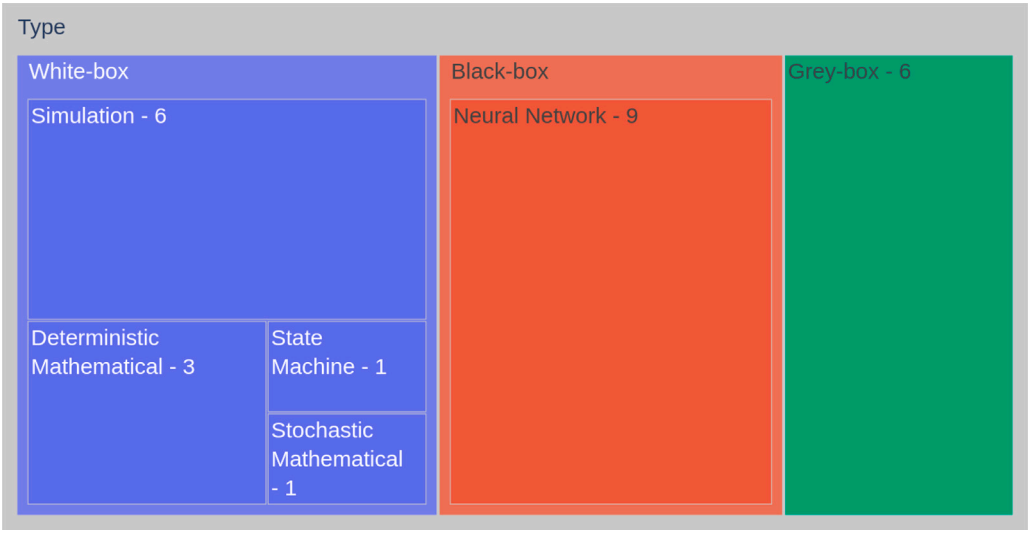


Fig. 6. White-box, black-box and grey-box modelling distribution.

other electronic specific parameters are required to determine failures and test failures as existing mathematical techniques for understanding their failures have been converted to work in a digital twin environment. Larger and more complex systems with more sensory inputs in this review tend to shy away from this method as it is hard to capture their inner workings through simple equations.

Stochastic mathematical Similar to deterministic mathematical techniques, stochastic mathematical techniques model a system using physics based equations but with the addition of random variables. In a modelling context, this approach is still classed as white-box as it is derived from physics-based processes [3].

Milton et al. [66] describes a mathematical based technique to developing a digital twin similar to that seen in Section 7.4.1. This digital twin, however, uses stochastic variables as part of the polynomial chaos expansion, which does not allow for inputs beyond this expansion to be mathematically determined, but instead inferred from a distribution. In a testing context, this approach would be classified as grey-box as the internal structure of the system is accessible, but the exact values required for white-box testing may not be.

State machine State machine based digital twins are those which replicate a system under test in the form of predetermined states and transitions which represent the system during operation. Eckhart et al. [53] uses a technique of multiple digital twins with their own states identical to those of their physical twin. The states are developed by users from a collection of current and historical data about the digital and physical assets, as defined in Section 7.3.1. A history of these states, as well as the transitions between them, can be used to identify behaviour which caused a critical state.

This technique does not use physics based equations [3,45] to outline behaviour but still provides a clear understanding of how data within the model will affect its state and change its output. Because of this, we define state machine based digital twins as white-box.

The states defined by digital twins are different to those found at the operation level of a cyber–physical system. These systems tend to have operational states to indicate actions, such as whether it is currently idle or processing. Digital-twin-based states provide contextual information, such as system-specific metadata [4], about the entire system as opposed to just an operational state.

White-box classifier White-box digital twins provide an insight into how the system is running or should be running. To achieve this, they must be compared to the physical asset using a classifier. This allows differences to be found and classifies failures. White-box classifiers provide data visibility and clarity during the failure classification process, allowing for more information during debugging [31].

Xiong et al. [67] proposes a classifier of an error vector based on the difference between vectors. All the outputs from the physical asset and a vector of the same outputs from the digital twin are compared. This simple classifier provides insight into which aspects of the physical asset differ from the digital twin and can therefore be investigated. Other studies [68,61,73] use classifiers for specific outputs from the physical asset and digital twin to find divergence with specified tolerances for each output. Both approaches provide full information visibility throughout their classification allowing for interpretation by an expert [31]. Although these two approaches appear very similar, input selection can provide insight into the testing technique used. This is further explored in Section 7.5.

7.4.2. Black-box

Black-box models are those which do not provide understandable representation of information transfer throughout the model and cannot be understood by experts in the domain [31]. Wagg et al. [3] defined black-box models in a digital twin context as those which are "derived entirely from measured data, with no assumed knowledge of the physics at all". The studies in this review contain testing techniques where both the digital twin models and any classifiers involved are designed in a black-box manner.

Neural network Neural network based digital twins are the most common technique found within studies in this review. Approaches in this review create the digital twin using a historical data set of the correct behaviour of the physical asset in order to mimic the correct behaviour of the system [69,71,74,52,59].

Neural networks are black-box as there is no easy way to determine how the inputs affect the outputs. This is because information visibility within the model is limited, and the model is derived purely from existing data [3,45]. A lack of information visibility and clarity of the derived model limits experts from determining the reasoning behind failures and erroneous behaviour [31].

It is not surprising that neural networks are the most common way to represent cyber–physical system behaviour. Historical data for existing cyber–physical systems can be easily obtained to provide highly accurate representations of correct behaviour. Using neural networks allows for a digital twin to act as an accurate replica of correct behaviour and is further explored by derived oracles in Section 7.3.

Table 7
Test case definition compared to test input types across the primary studies.

	General sensors	Specific sensors	Non sensor
Real-time monitoring	[52,74,72,53,70,75,54,55,61,56,57,60]	[64–66,59,69,68,62,71,63,73]	[76]
Predictive monitoring	[58,77]	–	–
Specified scenarios	–	[67]	–

7.4.3. Grey-box

There are techniques in the literature which use physics based digital twins, as explored in Section 7.4.1, but are classified in this review as grey-box. Grey-box digital twins combine both a physics based equation model with derived data based components which limit the visibility of information throughout the model [3]. By inferring data about the test failures using a black-box classification techniques, the overall internal clarity of the digital twin is reduced [32], but more powerful failure inference can be used.

Amini et al. [70] uses a neural network based classifier to determine failures within a physics based digital twin of wind turbines. Even though the digital twin is constructed as a simulation, test failure reasoning is inferred using a black-box technique, reducing information visibility throughout the model. This technique is further explored by using a classifier completely independent of the digital twin and cyber–physical system. Gao et al. [75] propose an anomaly detection classifier developed to work with real-time inputs from any digital and physical twins. Historical physical failure-free data as well as digital twin historical data are used to train an anomaly classifier linear model to provides insight into failure causes, but reduces information visibility throughout the model.

7.4.4. Summary

Digital-twin-based testing for cyber–physical systems uses a variety of white-box and black-box modelling techniques for designing digital twins. Neural networks are the most common implementation of digital twins, providing minimal data visibility within the model, but allowing the digital twin to more accurately represent its physical asset [3,31].

Specialised techniques for modelling digital twins, especially around electronic components, have been developed using mathematical representation. This conversion of existing component specific equations to digital twins has allowed for simpler modelling approaches where more complex approaches are not required.

Failure classifiers can be useful in providing a data derived approach to classifying erroneous behaviours, but their black-box nature can restrict information visibility within that section of the model.

As defined in Section 4, multiple different contexts were required for this review. These contrasting definitions made conducting this section of the review cumbersome. As digital twins are an emerging technology, this is not surprising but it would be useful to standardise definitions across the field to limit the confusion when conducting further reviews.

7.5. RQ4: How are test cases defined and how does this affect test inputs?

The testing procedures used to test cyber–physical systems are briefly covered in Section 7.2 to provide an overview of the different testing techniques used. This section explores how the test cases of these testing procedures are defined across different studies and how the inputs for those tests vary.

We split the test case types into three categories: “real-time monitoring”, “predictive monitoring” and “specified scenarios”. These categories allowed classification of each study and allowed for trends in active (predictive monitoring and specified scenarios) and passive (real-time monitoring) testing to be observed [28,47].

Table 7 outlines the distribution of different test case types compared to the input selection across the primary studies.

7.5.1. Real-time monitoring

Real-time monitoring between sensors on a cyber–physical system and the functional output of digital twin sensors allows for deviations between the two to be detected. This allows for any failures related to this to be found. This section investigates which sensors are monitored for the studies in this review.

Monitoring all sensors for a system in real-time is the most common technique found in this review. Xu et al. [74] and Grigoropoulos et al. [72] both compare each sensor in the physical asset with that of its digital twin to see if there are any deviations. This real-time monitoring allows for any erroneous behaviour to be identified during the run time of the system. This does, however, allow the system to enter erroneous states and does not test any specific scenarios which may be important.

Other studies, such as Peng et al. [64] and Milton et al. [66], provide sensor monitoring in real-time for only specific sensors in a system. This real-time approach provides the same benefits and drawbacks as the approach above, but instead may ignore deviation in unnecessary sensors to focus on those deemed more important. This approach may not capture all causes of a failure as only system specific sensors are monitored and therefore may miss other erroneous outputs not captured by the sensors under test.

Li et al. [76] ignores sensor outputs in favour of monitoring timing characteristics of a system compared to that of its digital twin. This ensures that failures can be detected when a system is running faster or slower than its digital twin as a deviation from the digital twin would imply that there is a failure in the physical asset.

7.5.2. Predictive monitoring

Predictive monitoring uses real-time monitoring of sensors within a system to predict if failures will occur. This is more of a proactive measure compared to the techniques outlined above as the system will be prevented from entering a state with a failure.

Xia et al. [58] and Cioroaica et al. [77] both use predictive monitoring to test that a user input on a physical asset can be safely performed by first simulating the end result on a digital twin before allowing the physical asset to follow. This approach does however require more computational power compared to other real-time sensor derivation methods and would introduce some latency into user inputs to the system. We found the fact that so few studies used predictive monitoring interesting as digital twins allow for increased predictive power.

7.5.3. Specified test cases

As well as monitoring a system for failures in real-time, some cyber–physical systems use specified test cases to ensure performance of the system in specific circumstances do or do not cause failures.

Xiong et al. [67] tests erroneous test scenarios which put the system under test into a state to detect failures which occur. This ensures that the digital twin is able to detect these states and report them accordingly. Due to the real-time nature of digital twins, very few studies used specified testing scenarios.

7.5.4. Summary

Section 3.2 provides classifications for digital twins as supervisory, interactive or predictive. The testing procedures found in this research question allow for an understanding of how digital twins of these classifications are used in testing. Most studies found interactive digital twins as they allowed for real-time monitoring to be achieved. These

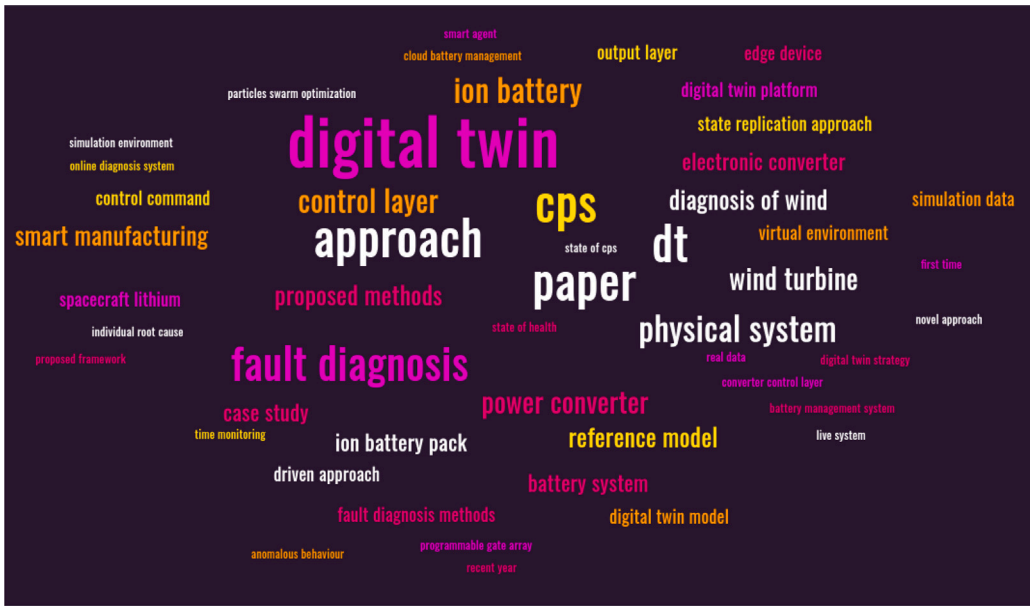


Fig. 7. Word cloud of primary study abstracts.

Table 8
Publication venues of the primary studies.

Publication venue	Count
Journal of Manufacturing Systems	2
Transactions on Power Electronics	2
Annual Conference of the PHM Society	1
Asia Conference on Power and Electrical Engineering	1
Automation in Construction	1
Conference on Software Testing, Verification and Validation	1
Cyber-Physical Systems Security and Privacy	1
IEEE Access	1
IEEE Systems Journal	1
Int. Conference on Advances in Artificial Intelligence	1
Int. Conference on Artificial Intelligence in Electronics Engineering	1
Int. Conference on Cyber-Physical Systems	1
Int. Conference on Industrial Internet	1
Int. Conference on Mechanical, Electric and Industrial Engineering	1
Int. Instrumentation and Measurement Technology Conference	1
Int. Symposium on Distributed Simulation and Real-Time Applications	1
Int. Symposium on Software Reliability Engineering Workshops	1
Journal of Energy Storage	1
Journal of Engineering	1
North American Manufacturing Research Conference	1
Sensors	1
Sustainability	1
Symposium on Information Control Problems in Manufacturing	1
Workshop on Cyber-Physical System Security	1

techniques change the physical asset's behaviour in response to the testing. A few studies, however, described predictive digital twins, further making use of the predictive capability of some digital twins. A future push towards more predictive digital twins would provide more predictive testing.

7.6. Publication trends

This section briefly summarises the publication trends of both publication venues for the primary studies in the area of cyber–physical system testing using digital twins, as well as the abstracts of the studies found within this review.

As is shown from Table 8, we outline the publication venues for the primary studies of this review. There are very few publication venues with multiple publications in the area of digital-twin-based cyber–physical system testing. This trend suggests that there is no specialised publication venue for this topic. This is expected for such an emerging technology, but the recent adoption shows that more specialised publication venues may be needed.

Fig. 7 shows a word cloud of the distribution of popularity of words and phrases within the abstracts of the accepted studies. It is interesting to see trends found from the results of the research questions found within this word cloud, such as the importance of failure diagnosis as well as the prevalence of digital twin testing for cyber–physical systems in manufacturing.

8. Discussion

In this section, we discuss the findings of this review, firstly discussing information arising from the methodology of the review, we then consider the results and the research questions, and finally reflect on the scope of this review.

8.1. Methodology discussion

This section explores the difficulties found during the creation of this review and possible solutions to those problems. We outline these issues in order to help future reviews and studies in this topic area.

8.1.1. Definitions of digital twins are fluid

How the technology of digital twins is defined is very different dependent on where you look. Section 3 explores several different definitions [4,8,21] as well as an applied definition in a digital twin framework [5]. Section 6.2.5 explored the expert opinion on the disparity of a definition and how it could negatively affect the use of grey literature in this review. As part of the inclusion and exclusion criterion IC3 and EC1 found in Section 6.2.4, this review screened studies based on their definition of a digital twin. This variety of definitions caused a large number of studies to be included in the initial search and then rejected during screening based on their definition of a digital twin. This is not a new problem as highlighted by Fuller et al. [6] stating that a standardised definition should be adopted.

A more concrete definition for digital twins would allow more reliable publication and identification of relevant materials. There have been attempts to standardise a definition; the British Standard for digital twins [41] aims to provide standardisation in digital twins, but some communities disagree with it due to its narrow scope focusing

primarily on manufacturing based systems. This review followed the definition set out by Eyre et al. [4] of a digital twin as it provides a relatively broad scope, yet does not include model based testing approaches such as hardware-in-the-loop testing. However, it could also be argued that imposing such a definition could stifle innovations in the form of new types of digital twins.

8.1.2. Digital twins are multidisciplinary

This review approached digital twins from a software testing perspective. In doing so, terms such as white-box, black-box and grey-box caused a lot of difficulty due to their uses in an engineering modelling context [3], as well as a software testing context [33]. This confusion in terminology caused challenges in both understanding the literature and writing this review.

In the future, publications could accommodate this by specifying context, especially when both engineering and software testing elements are present. This would reduce confusion and allow for a more accessible research field.

8.1.3. No concrete taxonomy for industries

The lack of an accepted taxonomy to cleanly split up different industries and sectors made answering RQ1 complicated as finding trends in different applications would have been very difficult without the ability to create groupings. We proposed the use of GICS [42] to solve this problem as it provided a multi-level definition for each industrial sector which allowed for simple classification of each study within this review.

Adoption of a universal taxonomy for aggregating applications into different industrial applications would allow for future reviews of technology with industrial application to proceed with an agreed standard.

8.2. Results discussion

This section discusses the results of the review. Observations from the results are highlighted and how this may affect the future directions of digital-twin-based cyber–physical system testing are investigated.

8.2.1. Industry 4.0 has affected adoption areas

The push for smart manufacturing as part of industry 4.0 has allowed for digital twins to be more prominent in manufacturing and other similar industrial applications [9]. This can be seen mirrored in the results to RQ1 with manufacturing, as well as other industrial applications, dominating the areas in which digital-twin-based testing is performed. This exemplifies the idea that digital twins are made to a high quality in established industries to ensure complex and modular systems, such as manufacturing plants, function correctly [43].

Industry 4.0 includes more applications than just manufacturing. Smart cities and autonomous vehicles are examples of systems with strong connections to industry 4.0, explored in Section 3.3, but were not uncovered as part of this review [22,23]. Smart cities are not covered by our definition of cyber–physical systems as they are defined as a system of systems. Therefore, smart cities were not within the scope of this review. Autonomous vehicles, however, could be defined as cyber–physical systems but were not captured by our search protocol. We assume this lack of literature is due to the early adoption of digital twins and that testing techniques will evolve as the subject matures. We believe that future work into the evolving implementation of testing procedures for applications such as this would be an intriguing avenue for future work.

8.2.2. Digital twins are becoming more accessible

Two studies [72,73], which fit into the information technology sector of the global industry classification standard [42], were found in the results for RQ1. These studies both used "off the shelf" modelling technologies to create their digital twins. The accessibility of digital twin building software, also mentioned in Section 6.2.5 with Amazon's twin maker [11], allows for easier adoption of the technology. As the technology becomes more accessible, this may require better standards, especially in safety-critical contexts, to ensure safety with physical assets.

Expansion of digital twins into new areas does, however, provide future challenges. The current British Standard for digital twins [41] focuses on digital twins in a manufacturing context and does not provide standardisation for implementations outside of this area. Future modifications to existing standards may be required to account for the adoption of digital twins into new areas.

8.2.3. Application specific testing techniques are beginning to emerge

Most of the testing techniques found in this review followed the technique of monitoring sensors and finding failures in their divergence. This trend is expected due to the infancy of using digital twins to test cyber–physical systems, especially in collaborative and modular systems [5] where a non-specialised approach may be adopted quickly. Results for both RQ1 and RQ4 uncovered that battery applications [64,65] of digital-twin-based testing followed the same, slightly more specialised, technique outlined in Sections 7.2.1 and 7.5.1 by modelling the degradation of the battery using existing battery mathematical techniques. This shows that specialised testing techniques are emerging.

8.2.4. Digital-twin-based conformance testing

Section 2.2 outlines four testing objectives present in testing cyber–physical systems. In this review, we found the majority of systems use conformance testing and that digital twins adapt the way in which conformance testing is performed. In digital-twin-based conformance testing, the oracle is defined as the digital twin, providing a dynamic representation for the expected behaviour of the cyber–physical system. This is different to traditional conformance testing which uses a static oracle. Due to this, digital-twin-based testing assumes that the digital twin is correct. We further explore the difficulties of a dynamic digital twin oracle in Section 8.2.6. It is interesting to see digital-twin-based conformance testing emerging as its own domain specific testing style in such a new area of application.

The studies in this review focus mainly on conformance testing with only Eckhart et al. [53] testing specifically in a security context. Milton et al. [66] briefly mention robustness and fragility testing, but more adoption of these testing objectives would be paramount for ensuring the rigorous testing of cyber–physical systems by digital twins. Work towards including these testing objectives would allow more comprehensive testing techniques within the stochastic environments of cyber–physical systems.

We were surprised to find that the majority of current testing techniques use failure monitoring for conformance as opposed to a variety of different techniques. The lack of diversity presents the infancy of the field as very few application specific testing techniques have been developed. As this area matures, we expect a broader range of testing techniques to be incorporated as well as failure monitoring.

8.2.5. Correctness of the digital twin is favoured over causality

The results found in RQ2 and RQ3 show a favouritism towards neural network based digital twins derived from correctly functioning historical data over simulation techniques derived from specification. Neural network based techniques allow for the digital twin to more correctly follow real life functionality which is important when finding failures in a running system [31]. This trend also fits an emerging technology as it allows for digital twins to be developed after the

fact, as opposed to during development of the cyber–physical system. Simulation based techniques, although they allow for visualisation, trade off accuracy for explainability and information visibility [3,31]. A lack of explainability in testing can cause more time debugging due to less information about the cause–effect relationship within the digital twin.

However, this trade-off of accuracy for explainability can be accounted for due to the use of neural network classifiers [48] to provide more insight into test failures. This can be seen reflected by the large number of grey-box techniques found in this review.

8.2.6. Oracle correctness is not tested

This review found that digital twins are used as the oracle when performing digital-twin-based testing against a cyber–physical system. Ensuring that the digital twin is working correctly in line with how the system should work is not mentioned in any studies within this review. The oracle correctness assumption states how oracles are assumed to be correctly functioning in software testing due to their derivation [49], although others argue that erroneous test oracles can occur and cause more problems as part of the debugging process [50]. Neural networks follow the oracle correctness assumption as they are assumed to be inherently correct in the way they are derived.

8.2.7. Digital twin prediction is underutilised

The results found in RQ4 show how real-time monitoring of when a cyber–physical system deviates from its digital twin is the most common form of testing found in this review. This approach allows for finding when a system enters an erroneous state but also still allows the system to first enter that state. The studies which use predictive monitoring in this review [58,77] do not allow for a system to enter an erroneous state but instead simulate ahead to prevent this.

Due to the predictive nature of some digital twins [9], this could be an interesting area for digital-twin-based testing to move into in the future to allow for a more proactive testing approach, especially in safety-critical roles where a system entering an erroneous state could cause harm.

8.2.8. Testing is mostly passive rather than active

RQ4 outlines the different types of test cases found in this review. Passive testing makes up the majority of testing techniques through real-time monitoring as opposed to active testing which can only be seen in one study through specific test scenarios [67].

Although testing is typically active with specific test cases and chosen inputs to reveal system behaviour [28], passive testing can also be used to monitor systems in real-time to ensure verification at run time [47]. As this subject matures, more specialised ways to incorporate active testing to complement the current passive testing would provide more reassurance in systems, especially in those which are safety critical.

8.2.9. Specific sensor testing is not always clear

RQ4 found that many studies within this review only used specific sensors on a system and its digital twin to test for failures. A large number of studies state that general sensors are used or do not specify which sensors are used during their testing. This lack of specification does not necessarily mean that all sensors are used in all cases, making it unclear whether specific sensors were used or not.

Grigoropoulos et al. [72] does not specify what data is relayed between the drone and its digital twin so it was classified as using general sensors in Section 7.5. In practice, only specific data would be shared between the two entities but it is unclear what that data is. This reduces the reproducibility of the studies as it does not comprehensively outline the testing technique used.

8.3. Review validity

In Section 6.2 we state that this review follows steps in accordance with Kitchenham et al. [36]. As part of this, it is important to explore both the internal and external validity of the review. Zhu et al. [51] takes this a step further by introducing construct validity as well as conclusion validity as well. This section will explore each of these validities and how this review may be threatened by each.

8.3.1. Construct validity

Construct validity focuses on how the review was constructed and specifically on the protocol used to obtain the studies. As explored in Section 6.2.1, the digital libraries used to find studies were chosen and tested to ensure they were inclusive of a set of test studies and then further expanded using snowballing, explored in Section 6.2.2, to find studies outside of the ones initially available. Section 6.2.2 also explores how the search string was generated and improved iteratively with the help of some automation techniques expressed in Section 6.2.3.

Although this process was thorough and ensured that a test set of studies could be found, it is still possible that studies were missed, especially due to the non-concrete definition of digital twins, which is discussed in Section 8.1.1.

8.3.2. Internal validity

Internal validity ensures that the cause–effect relationship of the methodology and results within the review are trustworthy. The search protocol, outlined in Section 6.2, was developed by one researcher which could be argued to introduce selection bias when executing the protocol. To reduce this bias, the protocol was reviewed by two co-authors who suggested changes and improvements to ensure there was no selection bias. Only when all authors were satisfied was the protocol executed.

8.3.3. External validity

External validity concerns how the results of the review are applicable to the subject area, which in this case is digital-twin-based testing for cyber–physical systems. Due to the lack of comprehensive definition of digital twins, as outlined in Section 3, it could be argued that this review is not valid for digital twins which do not fit the definition used by this review [4]. To minimise this threat to validity, a definition for digital twins in this review was outlined in Section 3 to ensure the domain in which this review is valid is more concrete.

8.3.4. Conclusion validity

Conclusion validity is defined by how reproducible the review and its results are. To ensure the results were reproducible, this review followed the steps outlined in Kitchenham et al. [36] to ensure the search methodology, outlined in Section 6.2, is both systematic and comprehensively documented. This validity is also vulnerable to the subjective definition of digital twins, outlined in Section 3, as the studies selected would differ based on this definition. To help combat this, a concrete definition of digital twins was outlined in Section 3.

9. Conclusion

This review produced a systematic search protocol which found 480 studies, manually reduced them to 26 through the use of abstract, introduction and full text screening. There are no other reviews which could be found on the topic of digital-twin-based testing for cyber–physical systems. We have provided an overview of the current testing landscape as well as outlined suggestions and future research areas in this field.

We found that the infancy of digital twins, especially in cyber–physical system testing, has caused issues in the writing styles of current literature. The lack of a concrete definition made determining

which literature is relevant to our review somewhat ambiguous. Ensuring a specification of data exchanged between the digital twin and its physical asset is presented guarantees reproducibility in future studies. The British standard for digital twins [41] provides some standardisation in this area, but is limited to manufacturing based applications. Future expansion of this standard into the emerging application areas found within this review could help its adoption. Further extension of this standard to include standardisation of oracle testing within the field would also be beneficial as we found a lack of oracle testing in this review. Increased confidence in failure detection due to better oracle testing allows for increased adoption and provides more predictive safety techniques to engineers.

Our research questions uncovered dominant and emerging areas that use digital-twin-based testing techniques. We outlined non-specialised testing techniques and the emergence of application-specific testing techniques. We found the use of specialised and derived oracles most common and that this affects the information visibility available for testing throughout the digital twin model. Real-time monitoring is currently the most common approach to testing using digital twins, with only very few studies using predictive digital twin capabilities in their testing.

As the subject matures, expanding testing to focus more towards active testing as well as expanding from primarily conformance testing could allow more testing potentials of digital twins to be adopted. Understanding and testing for the stochasticity of physical environments in which cyber–physical systems exist would provide a more rigorous form of testing to provide better confidence, especially in safety-critical systems. These safety-critical systems would also benefit from the predictive power of digital twins as a technology in their ability to proactively simulate ahead and prevent failures, saving money and preventing harm.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.infsof.2022.107145.

Data availability

Data has been made available on a public DOI (with respect to copyright limitations)

Acknowledgements

James A. Douthwaite is supported by the Assuring Autonomy International Programme (AAIP grant CSI: Cobot), a partnership between Lloyd's Register Foundation and the University of York, and from the UKRI project EP/V026747/1. David Wagg is funded by the UKRI, United Kingdom grant EP/R006768/1. Neil Walkinshaw and Rob M. Hierons are funded by the UKRI, United Kingdom CITCOM grant EP/T030526/1. Heirons is also funded by the UKRI, United Kingdom grants EP/R025134/2, EP/V026801/2. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY)[2] licence to any Author Accepted Manuscript version arising.

[2] Where permitted by UKRI a CC-BY-ND licence may be stated instead.

References

[1] R. Rajkumar, I. Lee, L. Sha, J. Stankovic, Cyber-physical systems: The next computing revolution, in: Design Automation Conference, 2010, pp. 731–736.
[2] X. Zhou, X. Gou, T. Huang, S. Yang, Review on testing of cyber physical systems: Methods and testbeds, IEEE Access 6 (2018) 1.
[3] D.J. Wagg, K. Worden, R.J. Barthorpe, P. Gardner, Digital twins: State-of-the-art and future directions for modeling and simulation in engineering dynamics applications, ASCE-ASME J. Risk Uncert Eng. Syst. B: Mech. Eng. 6 (3) (2020) 030901, http://dx.doi.org/10.1115/1.4046739, [Online]. Available:.
[4] J. Eyre, S. Hyde, D. Walker, S. Ojo, O. Hayes, R. Hartley, R. Scott, J. Bray, Untangling the requirements of a Digital Twin, Tech. Rep, Advanced Manufacturing Research Centre, 2021, [Online]. Available: https://www.amrc.co.uk/pages/digital-twin-report.
[5] J.A. Douthwaite, B. Lesage, M. Gleirscher, R. Calinescu, J.M. Aitken, R. Alexander, J. Law, A Modular Digital Twinning Framework for Safety Assurance of Collaborative Robotics, Tech. Rep, 2021, [Online]. Available: https://www.sheffield.ac.uk/sheffieldrobotics/about/csi-cobots/csi-project.
[6] A. Fuller, Z. Fan, C. Day, C. Barlow, Digital twin: Enabling technologies, challenges and open research, IEEE Access (8) (2020) 108952–108971.
[7] W. Kritzinger, M. Karner, G. Traar, J. Henjes, W. Sihn, Digital Twin in manufacturing: A categorical literature review and classification, IFAC-PapersOnLine 51 (11) (2018) 1016–1022.
[8] T. West, M. Blackburn, Is digital thread/digital twin affordable? A systemic assessment of the cost of DoD's latest manhattan project, Procedia Comput. Sci. 114 (2017) 47–56.
[9] Y. Hsu, J.-M. Chiu, J.S. Liu, Digital twins for industry 4.0 and beyond, in: 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2019, pp. 526–530.
[10] J. Wang, L. Ye, R.X. Gao, C. Li, L. Zhang, Digital Twin for rotating machinery fault diagnosis in smart manufacturing, Int. J. Prod. Res. 57 (12) (2019) 3920–3934, http://dx.doi.org/10.1080/00207543.2018.1552032, [Online]. Available:.
[11] A. Rasheed, Digital Twins on AWS: Unlocking Business Value and Outcomes, Tech. Rep., Amazon Web Services, 2022, [Online]. Available: https://aws.amazon.com/blogs/iot/digital-twins-on-aws-unlocking-business-value-and-outcomes/.
[12] A.R. Cavalli, T. Higashino, M. Núñez, A survey on formal active and passive testing with applications to the cloud, Ann. Telecommun. - Ann. TÉLÉCommun. 70 (3) (2015) 85–93, http://dx.doi.org/10.1007/s12243-015-0457-8, [Online]. Available:.
[13] J.L. Tekaat, H. Anacker, R. Dumitrescu, The paradigm of design thinking and systems engineering in the design of cyber-physical systems: A systematic literature review, in: 2021 IEEE International Symposium on Systems Engineering (ISSE), 2021, pp. 1–8.
[14] L. Piardi, P. Leitão, A.S. de Oliveira, Fault-tolerance in cyber-physical systems: Literature review and challenges, in: 2020 IEEE 18th International Conference on Industrial Informatics (INDIN), Vol. 1, 2020, pp. 29–34.
[15] B. Bagheri, S. Yang, H.A. Kao, J. Lee, Cyber-physical systems architecture for self-aware machines in industry 4.0 environment, 28, 2015, pp. 1622–1627.
[16] I.I. Standards, Systems and software engineering — Vocabulary: ISO/IEC/IEEE 24765:2017, 2017, Standard.
[17] H.L.S. Araujo, G. Carvalho, M. Mohaqeqi, M.R. Mousavi, A. Sampaio, Sound conformance testing for cyber-physical systems: Theory and implementation, Sci. Comput. Program. 162 (2018) 35–54, http://dx.doi.org/10.1016/j.scico.2017.07.002, [Online]. Available:.
[18] H. Abbas, B. Hoxha, G. Fainekos, K. Ueda, Robustness-guided temporal logic testing and verification for Stochastic Cyber-Physical Systems, in: The 4th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent, 2014, pp. 1–6.
[19] M.H. Cintuglu, O.A. Mohammed, K. Akkaya, A.S. Uluagac, A survey on smart grid cyber-physical system testbeds, IEEE Commun. Surv. Tutor. 19 (1) (2017) 446–464.
[20] D. Maclay, Simulation gets into the loop, IEE Rev. 43 (3) (1997) 109–112.
[21] L. Wright, S. Davidson, How to tell the difference between a model and a digital twin, Adv. Model. Simul. Eng. Sci. 7 (1) (2020) http://dx.doi.org/10.1186/s40323-020-00147-4, [Online]. Available:.
[22] R. Sell, A. Rassõlkin, R. Wang, T. Otto, Integration of autonomous vehicles and industry 4.0, Proc. Est. Acad. Sci. 68 (4) (2019) 389–394.
[23] D. Correia, L. Teixeira, J.L. Marques, Study and analysis of the relationship between smart cities and Industry 4.0: A systematic literature review, Int. J. Technol. Manag. Sustain. Dev. 21 (1) (2022) 37–66.
[24] M.S. Bonney, M. de Angelis, M. Dal Borgo, L. Andrade, S. Beregi, N. Jamia, D.J. Wagg, Development of a digital twin operational platform using Python Flask, Data-Centric Eng. 3 (2022) e1.

[25] M. Grieves, Product Lifecycle Management: Driving the Next Generation of Lean Thinking, McGraw-Hill Professional, 2005.
[26] K. Worden, E.J. Cross, P. Gardner, R.J. Barthorpe, D.J. Wagg, On digital twins, mirrors and virtualisations, in: Model Validation and Uncertainty Quantification, Volume 3, Springer, 2020, pp. 285–295.
[27] F. Milano, Power System Modelling and Scripting, Springer Science & Business Media, 2010.
[28] P.C. Jorgensen, Software Testing: A Craftsman's Approach, fourth ed., Auerbach Publications, 2013.
[29] E.T. Barr, M. Harman, P. McMinn, M. Shahbaz, S. Yoo, The oracle problem in software testing: A survey, IEEE Trans. Softw. Eng. 41 (5) (2015) 507–525.
[30] H. Agarwal, A. Sharma, A comprehensive survey of fault tolerance techniques in cloud computing, in: 2015 International Conference on Computing and Network Communications (CoCoNet), 2015, pp. 408–413.
[31] O. Loyola-Gonzalez, Black-box vs. White-Box: Understanding their advantages and weaknesses from a practical point of view, IEEE Access 7 (2019) 154096–154113.
[32] K. Worden, C. Wong, U. Parlitz, A. Hornstein, D. Engster, T. Tjahjowidodo, F. Al-Bender, D. Rizos, S. Fassois, Identification of pre-sliding and sliding friction dynamics: Grey box and black-box models, Mech. Syst. Signal Process. 21 (1) (2007) 514–534, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327005001408.
[33] M.E. Khan, F. Khan, A comparative study of white box, black box and grey box testing techniques, Int. J. Adv. Comput. Sci. Appl. 3 (6) (2012).
[34] D. Jones, C. Snider, A. Nassehi, J. Yon, B. Hicks, Characterising the Digital Twin: A systematic literature review, CIRP J. Manuf. Sci. Technol. (2020).
[35] C. Semeraro, M. Lezoche, H. Panetto, M. Dassisti, Digital twin paradigm: A systematic literature review, Comput. Ind. 130 (2021) 103469.
[36] B.A. Kitchenham, S. Charters, Guidelines for Performing Systematic Literature Reviews in Software Engineering, Tech. Rep. EBSE 2007-001, Keele University and Durham University Joint Report, 2007, [Online]. Available: https://www.elsevier.com/_data/promis_misc/525444systematicreviewsguide.pdf.
[37] K. Petersen, S. Vakkalanka, L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, Vol. 64, Elsevier, 2015, pp. 1–18.
[38] V. Garousi, M. Felderer, M.V. Mäntylä, Guidelines for including grey literature and conducting multivocal literature reviews in software engineering, Inf. Softw. Technol. 106 (2019) 101–121.
[39] G. Tsafnat, P. Glasziou, M.K. Choong, A. Dunn, F. Galgani, E. Coiera, Systematic review automation technologies, Syst. Rev. 3 (2014).
[40] M.M. Martín-Lopo, J. Boal, Á. Sánchez-Miralles, Transitioning from a meta-simulator to electrical applications: An architecture, Simul. Model. Pract. Theory 94 (2019) 177–198.
[41] B. Standards, Automation Systems and Integration. Digital Twin Framework for Manufacturing, BS ISO 23247:2021, Standard, 2021.
[42] M.S.C. International, Gics - global industry classification standard, 1999, https://www.msci.com/our-solutions/indexes/gics.
[43] W. ElMaraghy, H. ElMaraghy, T. Tomiyama, L. Monostori, Complexity in engineering design and manufacturing, CIRP Ann. 61 (2) (2012) 793–814, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0007850612002004.
[44] S. Ansari, A. Chandel, M. Tariq, A comprehensive review on power converters control and control strategies of AC/DC microgrid, IEEE Access 9 (2021) 17998–18015.
[45] K. Worden, R. Barthorpe, E. Cross, N. Dervilis, G. Holmes, G. Manson, T. Rogers, On evolutionary system identification with applications to nonlinear benchmarks, Mech. Syst. Signal Process. 112 (2018) 194–232, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327018301912.
[46] M. Koutsoubelias, N. Grigoropoulos, S. Lalis, A modular simulation environment for multiple UAVs with virtual WiFi and sensing capability, in: 2018 IEEE Sensors Applications Symposium (SAS), 2018, pp. 1–6.
[47] J. Wu, Y. Zhao, X. Yin, From active to passive: Progress in testing of internet routing protocols, in: M. Kim, B. Chin, S. Kang, D. Lee (Eds.), Formal Techniques for Networked and Distributed Systems, Springer US, Boston, MA, 2001, pp. 101–116.
[48] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019.
[49] T. Li, W. Liu, X. Guo, J. Wang, Software testing without the oracle correctness assumption, Front. Comput. Sci. 14 (3) (2020) 143203, http://dx.doi.org/10.1007/s11704-019-8434-4, [Online]. Available:.
[50] X. Guo, M. Zhou, X. Song, M. Gu, J. Sun, First, debug the test oracle, IEEE Trans. Softw. Eng. 41 (10) (2015) 986–1000.
[51] X. Zhou, Y. Jin, H. Zhang, S. Li, X. Huang, A map of threats to validity of systematic literature reviews in software engineering, in: 2016 23rd Asia-Pacific Software Engineering Conference (APSEC), 2016, pp. 153–160.

Primary Studies

[52] Y. Xu, Y. Sun, X. Liu, Y. Zheng, A digital-twin-assisted fault diagnosis using deep transfer learning, IEEE Access 7 (2019) 19990–19999.
[53] M. Eckhart, A. Ekelhart, A specification-based state replication approach for digital twins, in: CPS-SPC'18 : Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and Privacy : October 19, 2018, Toronto, on, Canada, Association for Computing Machinery, Inc, 2018, p. 108.
[54] M. Bevilacqua, E. Bottani, F.E. Ciarapica, F. Costantino, L.D. Donato, A. Ferraro, G. Mazzuto, A. Monterìù, G. Nardini, M. Ortenzi, M. Paroncini, M. Pirozzi, M. Prist, E. Quatrini, M. Tronci, G. Vignali, Digital twin reference model development to prevent operators' risk in process plants, Sustainability (Switzerland) 12 (2020).
[55] D. Gao, P. Liu, S. Jiang, X. Gao, K. Wang, A. Zhao, Y. Xue, Intelligent instrument fault diagnosis and prediction system based on digital twin technology, in: Journal of Physics: Conference Series, Vol. 1983, IOP Publishing Ltd, 2021.
[56] S. Kang, I. Chun, H.S. Kim, Design and implementation of runtime verification framework for cyber-physical production systems, J. Eng. (UK) 2019 (2019).
[57] E. Negri, S. Berardi, L. Fumagalli, M. Macchi, MES-integrated digital twin frameworks, J. Manuf. Syst. 56 (2020) 58–71.
[58] K. Xia, C. Sacco, M. Kirkpatrick, C. Saidy, L. Nguyen, A. Kircaliali, R. Harik, A digital twin to train deep reinforcement learning agent for smart manufacturing plants: Environment, interfaces and intelligence, J. Manuf. Syst. 58 (2021) 210–230.
[59] S. Yoginath, V. Tansakul, S. Chinthavali, C. Taylor, J. Hambrick, P. Irminger, K. Perumalla, On the effectiveness of recurrent neural networks for live modeling of cyber-physical systems, in: Proceedings - IEEE International Conference on Industrial Internet Cloud, ICII 2019, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 309–317.
[60] Q. Lu, X. Xie, A.K. Parlikad, J.M. Schooling, Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance, Autom. Constr. 118 (2020).
[61] M. Short, J. Twiddle, An industrial digitalization platform for condition monitoring and predictive maintenance of pumping equipment, Sensors (Switzerland) 19 (2019).
[62] M. Eckhart, A. Ekelhart, Towards security-aware virtual environments for digital twins, in: CPSS 2018 - Proceedings of the 4th ACM Workshop on Cyber-Physical System Security, Co-Located with ASIA CCS 2018, Association for Computing Machinery, Inc, 2018, pp. 61–72.
[63] C. Liu, A. Mauricio, J. Qi, D. Peng, K. Gryllias, Domain adaptation digital twin for rolling element bearing prognostics, in: "Proceedings of the Annual Conference of the PHM Society 2020", Vol. 12, 2020.
[64] Y.S. Yu Peng, Xu Zhang, D. Liu, A low cost flexible digital twin platform for spacecraft lithium-ion battery PackDegradation assessment, in: "2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)", IEEE Instrumentation and Measurement Society and Institute of Electrical and Electronics Engineers, 2019.
[65] W. Li, M. Rentemeister, J. Badeda, D. Jöst, D. Schulte, D.U. Sauer, Digital twin for battery systems: Cloud battery management system with online state-of-charge and state-of-health estimation, J. Energy Storage 30 (2020).
[66] M. Milton, C.O.D. La, H.L. Ginn, A. Benigni, Controller-embeddable probabilistic real-time digital twins for power electronic converter diagnostics, IEEE Trans. Power Electron. 35 (2020) 9852–9866.
[67] J. Xiong, H. Ye, W. Pei, K. Li, Y. Han, Real-time FPGA-digital twin monitoring and diagnostics for PET applications, in: Proceedings - 2021 6th Asia Conference on Power and Electrical Engineering, ACPEE 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 531–536.
[68] Y. Peng, S. Zhao, H. Wang, A digital twin based estimation method for health indicators of DC-DC converters, IEEE Trans. Power Electron. 36 (2021) 2105–2118.
[69] C.C. Peng, Y.H. Chen, Digital twins-based online monitoring of TFE-731 turbofan engine using fast orthogonal search, IEEE Syst. J. (2021).
[70] A. Amini, J. Kanfound, T.H. Gan, An ai driven real-time 3-D representation of an off-shore WT for fault diagnosis and monitoring, in: PervasiveHealth: Pervasive Computing Technologies for Healthcare, ICST, 2019, pp. 162–165.
[71] Q. Yu, Y. Huang, Y. Liu, S. Yu, S. Wang, Research on application of information model in wind turbine fault diagnosis, in: 2021 2nd International Conference on Artificial Intelligence in Electronics Engineering, in: AIEE 2021, Association for Computing Machinery, New York, NY, USA, 2021, pp. 67–74, http://dx.doi.org/10.1145/3460268.3460278, [Online]. Available:.
[72] N. Grigoropoulos, S. Lalis, Simulation and digital twin support for managed drone applications, in: 2020 IEEE/ACM 24th International Symposium on Distributed Simulation and Real Time Applications (DS-RT), 2020, pp. 1–8.
[73] C.M. Henson, N.I. Decker, Q. Huang, A digital twin strategy for major failure detection in fused deposition modeling processes, Procedia Manuf. 53 (2021) 359–367.
[74] Q. Xu, S. Ali, T. Yue, Digital twin-based anomaly detection in cyber-physical systems, in: Proceedings - 2021 IEEE 14th International Conference on Software Testing, Verification and Validation, ICST 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 205–216.

[75] C. Gao, H. Park, A. Easwaran, An anomaly detection framework for digital twin driven cyber-physical systems, in: ICCPS 2021 - Proceedings of the 2021 ACM/IEEE 12th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2021), Association for Computing Machinery, Inc, 2021, pp. 44–54.

[76] J. Li, O. Promyslova, V. Promyslov, K. Semenkov, The application of the hybrid digital twin for time characteristic assessment of the control system, IFAC-PapersOnLine 54 (2021) 965–970, [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2405896321008673.

[77] E. Cioroaica, F.D. Giandomenico, T. Kuhn, F. Lonetti, E. Marchetti, J. Jahic, F. Schnicke, Towards runtime monitoring for malicious behaviors detection in smart ecosystems, in: Proceedings - 2019 IEEE 30th International Symposium on Software Reliability Engineering Workshops, ISSREW 2019, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 200–203.