IntechOpen

IntechOpen Series
Artificial Intelligence, Volume 37

Anomaly Detection

Methods, Complexities and Applications

Edited by Miguel Delgado-Prieto

Anomaly Detection - Methods, Complexities and Applications

Edited by Miguel Delgado-Prieto

Published in London, United Kingdom

Anomaly Detection - Methods, Complexities and Applications
http://dx.doi.org/10.5772/intechopen.1004507
Edited by Miguel Delgado-Prieto

Contributors
Joan Valls Pérez, Mayra Ramírez Chávez, Miguel Delgado-Prieto and Luis Romeral Martínez, Susie Xi Rao, Jiawei Jiang, Zhichao Han and Hang Yin, Tee Hui Teo, Chiang Liang Kok, Chee Kit Ho, Xinlong Zhang, Jovan Bowen Heng and Guangming Ren, Ingo Elsen, Alexander Ferrein and Stefan Schiffer, Yash Patel

First published in London, United Kingdom, 2025 by IntechOpen
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, 167-169 Great Portland Street, London, W1W 5PF, United Kingdom

For EU product safety concerns: IN TECH d.o.o., Prolaz Marije Krucifikse Kozulić 3, 51000 Rijeka, Croatia, info@intechopen.com or visit our website at intechopen.com.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Anomaly Detection - Methods, Complexities and Applications
Edited by Miguel Delgado-Prieto
p. cm.

This title is part of the Artificial Intelligence Book Series, Volume 37
Topic: Applied Intelligence
Series Editor: Andries Engelbrecht
Topic Editor: Vladimir Robles-Bykbaev

Print ISBN 978-1-83634-359-2
Online ISBN 978-1-83634-358-5
eBook (PDF) ISBN 978-1-83634-360-8
ISSN 2633-1403

IntechOpen Book Series

Artificial Intelligence

Volume 37

Aims and Scope of the Series

Artificial Intelligence (AI) is a rapidly developing multidisciplinary research area that aims to solve increasingly complex problems. In today's highly integrated world, AI promises to become a robust and powerful means for obtaining solutions to previously unsolvable problems. This Series is intended for researchers and students alike interested in this fascinating field and its many applications.

Meet the Series Editor

Andries Engelbrecht received the Masters and Ph.D. degrees in Computer Science from the University of Stellenbosch, South Africa, in 1994 and 1999 respectively. He is currently appointed as the Voigt Chair in Data Science in the Department of Industrial Engineering, with a joint appointment as Professor in the Computer Science Division, Stellenbosch University. Prior to his appointment at Stellenbosch University, he has been at the University of Pretoria, Department of Computer Science (1998-2018), where he was appointed as South Africa Research Chair in Artifical Intelligence (2007-2018), the head of the Department of Computer Science (2008-2017), and Director of the Institute for Big Data and Data Science (2017-2018). In addition to a number of research articles, he has written two books, Computational Intelligence: An Introduction and Fundamentals of Computational Swarm Intelligence.

Meet the Volume Editor

Dr. Miguel Delgado-Prieto received the M. Sc. (2007) and Ph.D. (2012) in Electronics Engineering from the Universitat Politècnica de Catalunya (UPC), Barcelona, where he is an Assistant Professor in the Department of Automatic Control. His work spans technology transfer, project leadership, and doctoral supervision. His research focuses on fault-diagnosis algorithms and control techniques based on machine learning and deep learning, as well as real-time signal processing for industrial, energy, and transportation applications. He also serves as editor and contributor to several scientific books, facilitating the dissemination of knowledge among researchers and practitioners.

Contents

Preface

Anomaly detection, the discipline concerned with identifying rare, atypical, or unexpected patterns in data, has assumed strategic importance across various industries, including manufacturing, power systems, electronic commerce, and even astrophysical research. As contemporary infrastructures grow in complexity and interconnectivity, the timely recognition of deviations becomes indispensable to operational safety, resilience, and efficiency.

Anomaly Detection – Methods, Complexities and Applications gathers recent work that advances this agenda on two fronts. The first section reviews algorithms and methodological frameworks, while the second illustrates how those ideas perform when confronted with the constraints of real-world data and deployment environments.

The opening section, "Anomaly Detection Techniques and Algorithms", lays the methodological groundwork. Chapter 1, "Recent Progress of Anomaly Detection in Energy Applications: A Systematic Literature Review", examines fifty-two studies on renewable generation, building consumption and energy storage, tracing the rise of ensemble and deep-learning methods while exposing gaps in critical-infrastructure and electric-vehicle-charging contexts. Chapter 2, "Fraud Detection in E-Commerce: A Systematic Review of Transaction Risk Prevention", maps the shift from rule-based screens to multimodal deep-learning pipelines, highlighting explainability, human-in-the-loop practices, online scalability and emergent themes such as federated and adversarial learning.

The second section, "Anomaly Detection in Practice", follows with three application-driven studies. Chapter 3, "Supervised Anomaly Detection with Attention", presents a temporal-convolutional architecture that leverages labelled data and attention cues to refine machinery-fault diagnosis, remaining-useful-life prediction, and real-time embedded monitoring of electric-fan drives. Chapter 4, "Anomaly Detection in Metal-Textile Industries", introduces a student–teacher feature-pyramid network that automates most optical inspections of gas-filter fabrics and routes only ambiguous cases to human inspectors, securing zero-defect quality at line speed. Chapter 5, "Predicting Exoplanets Habitability: Metrics and Models", applies a variational auto-encoder to imbalanced stellar catalogues, isolating the rare orbital and atmospheric configurations most conducive to life and refining conventional habitability indices.

This volume represents the coordinated effort of many hands. I am grateful to the authors for their scholarly diligence, to the anonymous reviewers for their thoughtful criticism, and to the editorial assistants and the IntechOpen production team for steering the project to completion with characteristic professionalism. Their collective work has shaped a coherent narrative out of varied perspectives.

I hope that the chapters which follow will not only serve as a reference but will also spark new collaborations and research questions, ultimately refining the methods we rely on to keep increasingly complex systems secure, reliable, and insightful.

Miguel Delgado-Prieto PhD.
Automatic Control Department,
Polytechnic University of Catalonia,
Barcelona, Spain

Section 1

Anomaly Detection Techniques and Algorithms

Chapter 1

Recent Progress of Anomaly Detection in Energy Applications: A Systematic Literature Review

Joan Valls Pérez, Mayra Ramírez Chávez, Miguel Delgado-Prieto and Luis Romeral Martínez

Abstract

Over the past few years, the anomaly detection problem has been intensively researched within different areas and applications. From a data-based analysis point of view, anomalies can be defined as data points that represent non-typical events, that is, abnormalities, with respect to the rest of the considered observations. The importance of anomaly detection relies on the fact that abnormal data highlights potentially undesirable situations in regard to the underlying physical phenomena under observation, which can have severe consequences for human beings, nature, infrastructures or information. This review article intends to provide a comprehensive overview of recent work on anomaly detection in a critical sector that is experiencing a deep digital transformation: the energy sector. With that, 52 articles have been reviewed, most of which focus on renewable energy generation, building energy consumption and energy storage. Interestingly, artificial intelligence-based approaches are found in ensemble schemes, where different models are combined for the maximization of the anomaly detection performance, oftentimes including deep learning (DL) models. However, under-represented trends and knowledge gaps are also identified, underscoring the lack of articles referring to specific energy application domains, such as critical infrastructures and electric vehicle (EV) charging infrastructure, and open issues for specific methodologies, such as explainability and applicability for deep learning anomaly detection solutions. Further, emerging concepts are highlighted and future research directions are identified.

Keywords: anomaly detection, outlier detection, fault detection, energy applications, renewable energy, artificial intelligence

1. Introduction

Anomaly detection involves identifying patterns within data that are different from expected behavior. These irregular patterns are often referred to as anomalies, outliers, novelties, discordant observations, exceptions, aberrations, peculiarities or contaminants, depending on the application domain [1]. These terms are largely interconnected and often overlap, though they carry subtle distinctions in meaning.

The literature offers a variety of definitions for abnormal observations or outliers. Some describe them as data points that deviate significantly from the rest of the sample [2] to the extent that they raise suspicion of having been produced by a different underlying process [3]. Similarly, more recent definitions [4, 5] refer to the concept of local density or relative density or even mention clusters of data.

A key aspect of anomaly detection lies in understanding the type of anomaly being identified. There is a general consensus in the literature on categorizing anomalies into three main types: point anomalies, collective anomalies and contextual anomalies [1, 6, 7]. A point anomaly refers to an individual data point that significantly differs from the rest of the dataset. When such a deviation is only considered anomalous within a specific context, it is termed a contextual anomaly. If a group of related data points is collectively unusual compared to the overall dataset, it is identified as a collective anomaly [1]. Other references [8, 9] introduce the concept of continuous anomaly, which closely aligns with the definition of collective anomaly. Meanwhile, Samariya and Thakkar [10] expand this classification by introducing a fourth category, termed group anomaly, with a slightly different organizational structure compared to the framework proposed by Chandola et al. [1].

While point anomalies are generally considered alarm triggers that must be addressed based on predefined knowledge, such as the ones caused by a momentary malfunction in the data acquisition, transmission or processing chain, collective and contextual anomalies typically uncover a more complex scenario. These types of anomalies often lead to the discovery of new behavioral patterns within the system under analysis.

Thus, anomaly detection application domains range from industry to finance and cybersecurity, as well as e-commerce and logistics. However, one of the sectors that is currently attracting significant attention, due to its ongoing transition to a digital framework, is the energy sector, particularly regarding energy generation and distribution, monitoring and management of smart grids.

The proliferation of intelligent decision-making procedures at different levels of smart asset management (i.e., primary, secondary and tertiary), along with the availability of high computing capacities both at the edge and in the cloud, is leading to the development of multiple data-driven solutions. These solutions enable, among other functionalities, anomaly detection related to system operation. Thus, due to its current trend of digitalization and its various levels of application, ranging from individual measurement devices to energy grid management, anomaly detection is a key aspect in the energy sector, which acts as the main motivation of the present study.

Consequently, conducting a systematic literature review (SLR) was deemed valuable to offer an overview of recent developments in anomaly detection research. Thousands of relevant publications were systematically screened from two major online databases, ultimately resulting in a curated set of 52 papers related to anomalies. Information from these studies was extracted and synthesized to address the authors' research questions (RQs). Ultimately, the main contributions of this work are the responses to the research questions and their sub-questions, presented in Section 4.

- RQ1: What recent peer-reviewed studies apply anomaly detection to energy systems, and what are their key methods, data types, applications and evaluation metrics?

- RQ2: What are the open issues of anomaly detection research?

The main contribution of this work lies in a comprehensive synthesis and interpretation of the current state of the art regarding anomaly detection in the energy sector. Thus, a systematic literature review (SLR) has been considered, which includes search, selection, classification and analysis of the most recent research studies in this field.

This approach not only provides an in-depth understanding of advancements in the specific domain, which are highlighted throughout this work, but also facilitates the identification of trends and developments, current challenges and research opportunities, as well as the formulation of taxonomies and theoretical frameworks that enable comparative analysis of approaches and critical evaluation of research directions.

The originality of this study includes consideration of recent research works, its specific emphasis on the energy sector, and its discussion structured around three main axes: applications, methodologies and algorithms.

The remainder of this chapter is structured as follows. Section 2 provides some background in the form of a taxonomy focused on anomaly detection in energy applications. Then, in Section 3, the chosen approach to conduct this systematic literature review is presented. Key results and findings are described in Section 4, with special attention to open issues and future research directions. Finally, Section 5 concludes the chapter, highlighting the major findings and most relevant concepts.

2. Background

Anomaly detection approximations have been developed within diverse research areas, and many anomaly detection taxonomies have been published over time, some of which focus on a specific type of algorithm (e.g., deep learning, generative adversarial networks (GANs), transient wave-based methods) [6, 11–13] or a particular application domain (e.g., network communications, internet of things, time series data, financial data) [14–16], while others are more generic [1].

Longstanding publications [1, 17] refer to seven different anomaly detection categories, characterized by particularities related to the input data and whether they are labeled, their type of anomaly among those described in Section 1, and the type of output data, such as score or label. Specifically, statistical, nearest neighbor-based, isolation-based, spectral techniques, information-theoretic techniques, classification-based and clustering-based. Although some previous references [18–20] also mention link-based and model-based approaches, this nomenclature has not been coined, which could relate to the fact that words such as *link* and *model* have become generic in the literature.

Statistical anomaly detection is the earliest work that has been proposed for anomaly detection [10], as it relies on well-known statistical models and statistical tests. In this approach, a statistical model is fitted to the given data, and then an inference test is applied to determine if an unseen instance belongs to the fitted model, which would be declared normal behavior or an anomaly otherwise [1]. Anomaly detection algorithms based on statistical models can be classified into parametric and non-parametric according to their assumption of the knowledge of the underlying distribution [21]. Parametric algorithms, such as the maximum normed residual (MNR) test or Dixon test, assume that the underlying distribution is known, whereas non-parametric algorithms, such as histograms or kernel density estimation (KDE), do not assume any prior knowledge of the data distribution. Statistical techniques' strengths are their simplicity and intuitiveness, but they tend to struggle with

high-dimensional data, which is very common in many applications. The concept known as the curse of dimensionality, first introduced by Bellman [22], describes the issue of data sparsity that arises as the number of input dimensions increases. This sparsity can hinder the effectiveness of anomaly detection methods, as the presence of irrelevant or redundant attributes may mask or distort the abnormal characteristics of the data [23]. This effect, although it has an impact on any data-based analysis approach, poses particular challenges for statistical and statistical-based methods. For instance, distance-based measures become less reliable in high-dimensional spaces, as data points tend to appear nearly equidistant from one another, a direct consequence of the curse of dimensionality [24].

Nearest neighbor-based methods operate under the assumption that normal instances tend to reside in dense regions of the data space, while anomalies are typically isolated and distant from their nearest neighbors. These techniques rely on a measure of distance or similarity between data points, which can be calculated in various ways. Broadly, nearest neighbor approaches can be classified into two categories: distance-based and density-based methods. Distance-based methods, such as k-nearest neighbor (kNN), evaluate anomalies by measuring the distance from a data point to its kth nearest neighbor. In contrast, density-based approaches, such as the local outlier factor (LOF) or connectivity-based outlier factor (COF), assess anomalies by comparing the local density of each instance to that of its neighbors to compute an anomaly score [1]. Both approaches have been deeply researched, and several improvements have been proposed over the years. Recent works grant an independent category to each of these two approaches [10], whereas others refer to nearest neighbor anomaly detection as neighbor-based [25]. Performance of statistical and nearest neighbor approaches decreases with high-dimensional datasets with varying sample densities, which fueled the inception of isolation-based anomaly detection.

Isolation-based anomaly detection represents a fundamentally distinct methodology, as it does not rely on traditional distance or density calculations. Instead, it leverages two key characteristics of anomalies: their rarity and their distinct attribute values compared to normal data points. The earliest implementation of this approach introduced the use of a binary tree structure known as an isolation tree (*iTree*) and a combination of them in the form of an isolation forest (*iForest*), which are very simple mechanisms that are effective and efficient in detecting anomalies [17]. Building upon these mechanisms, many extensions and variations have been proposed, such as separate clustered isolation forest (*SciForest*) or local sensitive hashing isolation forest (*LSHiForest*), to address different challenges and improve performance [26]. Isolation measures improve handling datasets with regions of different densities, which is one of the main weaknesses of distance-based and density-based methods. Some recent works [26] refer to isolation mechanisms and their application in other anomaly detection-related tasks, such as clustering and classification, whereas others [10] refer to isolation as an individual anomaly detection approach.

Spectral methods aim to approximate the dataset by identifying a combination of attributes that capture most of its variance, operating under the assumption that the data can be effectively represented in a lower-dimensional subspace where normal and anomalous instances become more distinguishable [1]. In some recent works, spectral techniques are primarily principal-component-analysis-based techniques [27, 28], which are in line with the aforementioned assumption. However, other recent research [10, 25] refers to subspace-based anomaly detection under the same assumption, thus avoiding decomposition-focused algorithms and presenting algorithms that combine selection space and anomaly detection, such as subspace outlier degrees (SOD) and LOF.

Information-theoretic approaches examine the informational characteristics of datasets through metrics such as Kolmogorov. Complexity, entropy and relative entropy are based on the premise that anomalies introduce disruptions in the overall information structure of the dataset [1]. Although some very recent works [29] consider an information-theoretic characterization for anomaly detection in data compression environments, recent anomaly detection surveys [16, 24, 25, 28, 30] no longer refer to information-theoretic techniques as a specific approach for anomaly detection.

Classification-based and clustering-based terminology is frequently used to distinguish between supervised and unsupervised learning settings, respectively. In classification-based anomaly detection, for example, a model is built using a labeled dataset and subsequently employed to assign new instances to predefined classes based on the learned patterns. Thus, classification-based approaches consist of solving supervised problems, that is, with a priori knowledge (i.e., historical labeled data), by means of neural network (NN) models, support vector machines (SVMs) and rule-based schemes. On the other hand, clustering-based anomaly detection approaches consist of solving unsupervised problems or semi-supervised problems, that is, with unlabeled or partially labeled data, by means of clustering algorithms such as self-organizing maps (SOMs), k-means clustering and expectation maximization (EM) [1].

Over the last two decades, the proliferation of data processing techniques based on artificial intelligence, which range from supervised, semi-supervised and unsupervised, has promoted the inception of new approaches, which have had an effect on the already existing categories found in the literature. Specifically, recent work [28] mentions classification-based approaches and clustering approaches as two of the main categories, which also include artificial intelligence-based approaches, while other recent authors [30] refer to deep learning-based approaches as a main category. As the datasets gradually become larger and more complex, more deep learning (DL) models have been proposed to perform anomaly detection tasks, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), auto-encoders (AEs) and generative adversarial networks (GANs) [12]. Other recent works [10, 25, 28] mention machine learning (ML) when referring to an approximation that is not found in longstanding publications, namely ensemble-based anomaly detection or hybrid anomaly detection, which consists of simultaneously combining different learning techniques or even multiple subspaces, where the potential anomalies are derived by ensemble techniques.

Despite of the certain categories which are considered in recent literature reviews on anomaly detection, there is no generalized consensus regarding the main approaches. In some cases, conceptual synonymy can be observed among terms used to define specific categories, as seen in the case of ensemble or hybrid approaches. While longstanding publications refer to subcategories at a second level of taxonomy [1], more recent studies tend to limit themselves to a single level to differentiate the main approaches, using the second level to reflect the relationship between techniques associated with each approach. Thus, categories appear to serve as a general framework within which the included techniques share theories and analytical foundations. However, some categories may exhibit set intersections, such as spectral and statistical approaches, for instance, in the case of principal component analysis (PCA).

3. Methodology

An SLR was carried out by referring to other pertinent guidelines and studies cited in [31–33]. Guided by the research questions outlined in Section 3.1, relevant search

terms were defined (Section 3.2), and a search strategy was developed (Section 3.4). This strategy incorporates inclusion and exclusion criteria detailed in Section 3.3 to minimize potential bias during the selection process. The main results from the selected studies are presented in Section 4.

3.1 Research questions

To answer the general research questions raised in Section 1, we detail them into sub-questions, as research questions of an SLR are usually generic and related to research trends. To be more specific on what characteristics of the most recent anomaly detection studies are to be examined, RQ1 is divided into five sub-questions.

As discussed in Section 2, the overview of anomaly detection approximations has evolved over the years, with new approximations gaining interest while others become less mentioned in the literature. It is important to understand which key methods are being considered in the most recent publications. Firstly, from the approximation point of view. RQ 1.1—*Which anomaly detection approximations were most researched?* However, as it was noted in Section 2, although different authors converge when referring to specific anomaly detection approaches, there are also significant differences regarding the organization of the main approaches and regarding the specific algorithms for each category. Secondly, it is important to investigate from the algorithms and techniques point of view, which will also provide insight about the popularity of anomaly detection-based applications of artificial intelligence algorithms. RQ 1.2—*Which algorithms and techniques were implemented?*

Anomaly detection is used in several application domains, such as intrusion detection systems, fraud detection and fault detection, where anomalies can relate to people, systems and processes, with very different meanings. From an anomaly detection in energy applications point of view, which is closely related to industrial devices and communication systems, anomaly detection algorithms and techniques can be implemented in different hierarchical layers, that is, component, device, system, process and plant. RQ 1.3—*In which hierarchical level were anomaly detection techniques implemented?*

Availability of datasets and data spaces is essential for developing anomaly detection applications that are relatable to real-world scenarios. In that sense, it is paramount to characterize the nature of data used for the validation of algorithms and techniques in anomaly detection applications. RQ1.4—*What were the main characteristics of the selected datasets? Were the datasets publicly available?*

Finally, as in many other knowledge fields, evaluation metrics are a key aspect of the validation of anomaly detection solutions, as they enable benchmarking the performance of innovative anomaly detection schemes against the performance achieved by ones that have already been developed and consolidated in the literature. Therefore, it is a relevant aspect to be investigated. RQ1.5—*Which are the reference evaluation metrics that have been used more recently?*

3.2 Search string

Based on the research questions, search terms were determined and organized into three distinct categories. As discussed in Section 1, anomaly detection, novelty detection and outlier detection are sometimes used indistinctly. Therefore, the first group contains the keywords:

- Group 1: (“anomaly detection” OR “novelty detection” OR “outlier detection”).

As presented in Section 2, specific anomaly detection main categories are found in the literature, although there is no consensus regarding the organization of the categories. The second group is related to the approximations that are included in the present study, represented by the following keywords:

- Group 2: ("statistics" OR "nearest neighbor" OR "distance" OR "density" OR "isolation" OR "spectral" OR "subspace" OR "information theory" OR "classification" OR "survey" OR "ensemble" OR "hybrid").

Finally, regarding the outcome of the contribution, which is related to the goals of the anomaly detection publications, represented by the following keywords:

- Group 3: ("architecture" OR "design" OR "verification" OR "validation" OR "test" OR "analysis").

To design the search string, the conjunction of the group terms above was used, that is, Group 1 AND Group 2 AND Group 3. Thus, the string designed for the search was the input for the procedure, which will be described in Section 3.4.

3.3 Inclusion and exclusion

The aim of this SLR was to identify and classify papers related to anomaly detection approaches for energy applications. The inclusion criteria (IC) were:

- (IC1) The paper must refer to an anomaly detection application.
- (IC2) The paper must have an energy context.
- (IC3) The paper must be published between 2020 and 2025.

Papers were excluded if they met any of the following exclusion criteria (EC), which were:

- (EC1) Studies that do not focus on energy are omitted.
- (EC2) Gray literature and papers written in languages that are not English are discarded.
- (EC3) Studies and documents that are not peer-reviewed are also excluded, as well as books, theses and dissertations.
- (EC4) Papers published in journals with SCImago Journal Rank (SJR) [34] below 1.0 are excluded.

3.4 Search strategy and selection process

The search strategy developed consists of four main phases and has been designed around two reference databases, aiming to maximize the chances of retrieving anomaly detection papers from different research communities. **Figure 1** depicts the search strategy, providing the number of studies that resulted from conducting each phase.

Particularly, the search string, which was introduced in Section 3.2, was used, with adaptation if necessary, on two online databases: ACM DL [35] and Scopus [36]. The

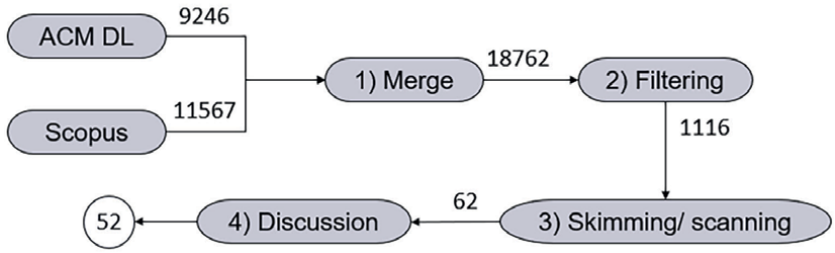

Figure 1.
Search strategy steps.

main reason for using these databases is that they are well-known within the research community, easy and intuitive to use and provide access to a vast number of scientific papers, which can be exported easily into reference managers.

- Step 1. The results obtained from the various search engines were combined to remove any duplicate entries.
- Step 2. Books, white papers and conference articles were manually removed. Journal articles published earlier than 2020 were removed. From the resulting set of journal articles, the ones published in academic journals with SJR below 1.0 were excluded.
- Step 3. A large portion of the candidate papers was excluded based on the established inclusion and exclusion criteria. The selection process involved multiple levels of review, including evaluation of the title and abstract and a quick examination of the main content. Papers that raised any uncertainty during the process were retained for more detailed assessment at a later stage.
- Step 4. Papers falling near the inclusion threshold were reviewed collaboratively by the authors to make final decisions regarding their inclusion or exclusion. Ultimately, a total of 52 papers were selected for the final corpus.

4. Results and discussions

In this section, the outcomes of the review are addressed. First, the most significant publication trends are presented graphically, aiming to provide a simple yet illustrative overview of the selected publications. Then, results of the research questions and sub-questions are provided in an organized manner, presenting the most relevant aspects and revealing significant insights. Finally, based on the presented findings for the research questions, open issues for anomaly detection in energy sector applications are pointed out, and research directions are proposed.

4.1 Publication topics

The analysis of the distribution of the selected anomaly detection studies can be investigated by focusing on different dimensions, which provide interesting insights. First, the number of articles published for each year is illustrated in **Figure 2**, which shows a steady growth in the number of anomaly detection publications for energy

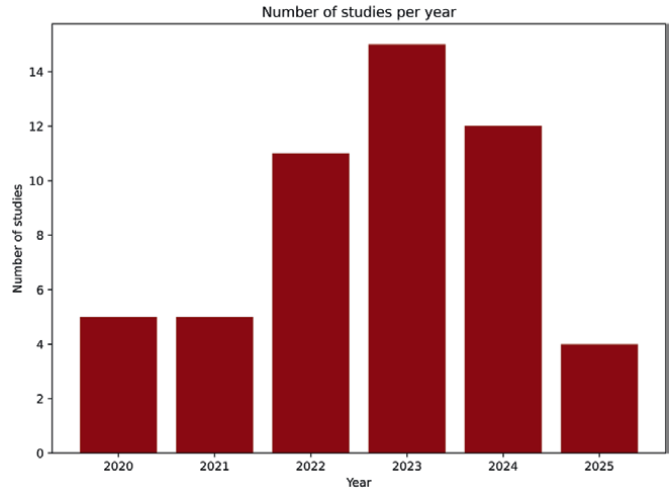

Figure 2.
Distribution of the selected articles per year.

applications for the last 5 years. In that sense, 2025 should be ignored when looking for a year-to-year trend, as the time of writing was the first quarter of 2025. Thus, the increase in published papers is believed to support that anomaly detection in energy applications is becoming increasingly relevant.

Now, the distribution of studies per journal of publication is investigated to capture, at a glance, which areas were of interest within energy-focused anomaly detection applications. The bar chart in **Figure 3** illustrates this, highlighting the interest

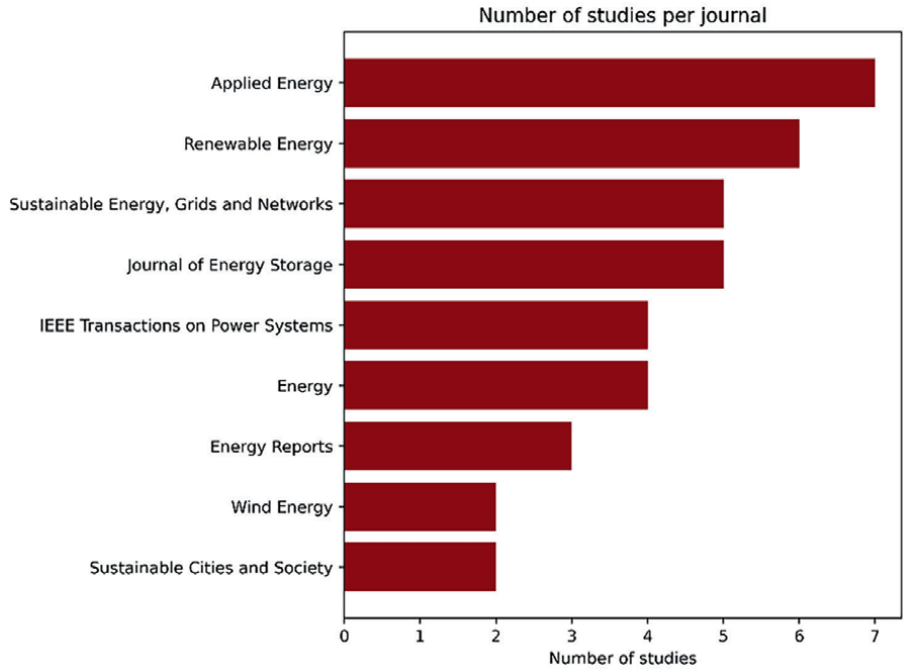

Figure 3.
Distribution of the selected articles per journal.

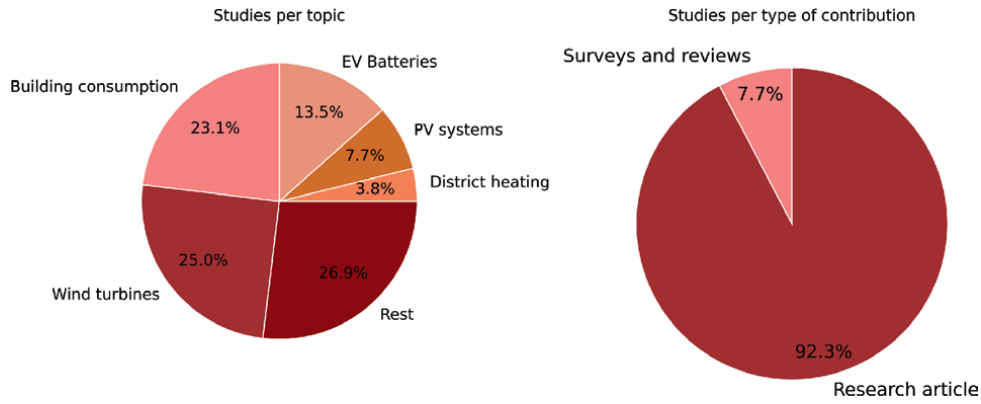

Figure 4.
Distribution of the selected articles per topic (left) and per type of contribution (right).

in research on anomaly detection applications related to renewable energy generation systems, energy storage systems, power systems and smart grids.

From the publication type perspective, the pie chart in **Figure 4** shows that only 7.7% (4 out of 52) of the selected studies are reviews or surveys. These publications address three of the four topics that receive the most attention in the remaining articles that comprise the sample selected in the present study, which are building electric consumption data, district heating and electric vehicle batteries.

Furthermore, regarding publication topics, there is a significant number of publications that focus on renewable energy sources, which is also depicted in **Figure 4**. Remarkably, a quarter (i.e., 25%) of the publications (13 out of 52) refer specifically to anomaly detection applications on wind turbines, namely, condition monitoring, fault detection and power curve modeling. Following is a brief introduction of the other topics that receive attention, underscoring their importance:

- The availability of energy consumption data has increased with the massive deployment of specific metering solutions (i.e., smart meters or smart sensors) or smart devices that include metering capabilities (i.e., smart inverters). With that, different stakeholders in the electrical system, end-users, energy producers or utility companies, are able to identify deviations in consumption patterns that can be investigated for different purposes. From the anomaly detection perspective, energy consumption data allows for applications such as energy saving, energy theft, theft attack detection, occupancy detection, home elderly monitoring and fault detection of the energy systems. In one of the selected reviews [37], artificial intelligence techniques for anomaly detection are thoroughly discussed, with consideration not only of the most relevant techniques and algorithms but also of computing environments and application domains.

- Efficient heat distribution in urban areas is a key application for energy management and efficiency-improving research, which is one of many application domains for anomaly detection. Specifically, district heating substations (DHS) are drawing interest as they are being increasingly used for affordable, low-carbon heat supply, which can be directly used by customers [38]. In short, the function of DHS is to hydraulically separate the water in the district heating circuit from that in the end-user installation. For that, fault detection is the most

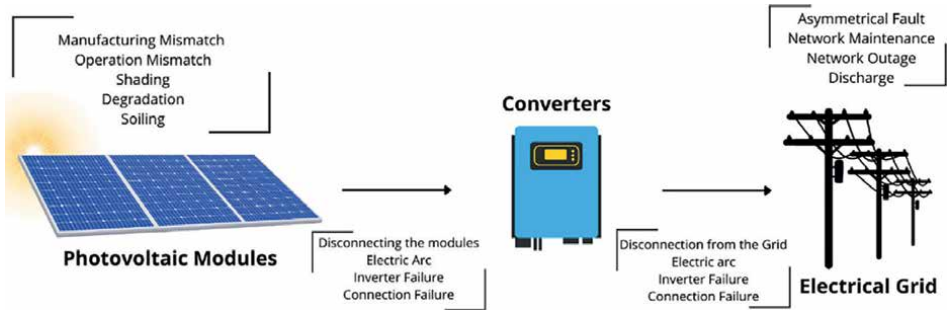

Figure 5.
Various problems that can cause anomalies in PV systems [40].

concerning anomaly detection application for DHS, as faulty or worn components can cause water leakages, which would be a waste of resources and could potentially damage nearby components.

- Aside from wind energy, which is the most popular topic within the selected articles, solar photovoltaic energy (PV) has become a widespread renewable energy technology, with regulators incentivizing the installation of industrial, tertiary and residential buildings, as well as with the decrease in the price of the required components for the installation, mainly the PV panels. From the anomaly detection perspective, detection and classification of anomalies in these systems is critical for ensuring a sustainable expansion of this technology, as well as the reduction of energy costs for the end-user [39]. **Figure 5** depicts some of the problems that can cause anomalies in PV systems.

- Electric traction solutions are booming due to the increasing pressure on vehicle manufacturers in the form of restrictions related to emissions from internal combustion engine vehicles. In this sense, all elements, components and subsystems found in electric powertrains of electric vehicles (EVs) are receiving attention from the perspective of design and operation optimization. Battery safety and reliability, in particular, are of great interest, given their high manufacturing cost caused by the complexity of the manufacturing process and the scarcity of the required materials. Thus, research into the development of supervision and monitoring systems, typically battery management systems (BMSs), with anomaly detection functionalities, is critical to safely maximize the life of these elements. In particular, lithium-ion batteries, which are the most common in mass-produced electric vehicles, present anomalies related to energy efficiency and safety [41].

Finally, it is worth noting that terminology-wise, most studies refer to *anomaly detection* and *outlier detection*, which suggests that *novelty detection*, which was included in the search string presented in Section 3, has not been coined in anomaly detection in energy applications.

4.2 The characteristics of anomaly detection studies

This section describes the main results to answer RQ1 and its sub-questions. Each subsection focuses on one of the sub-questions, detailing the most interesting aspects

of the selected papers and illustrating the differences in distribution among the four main topics by means of comparative tables.

4.2.1 Anomaly detection approaches

Anomaly detection approaches that are referred to in the review articles are subject to different notions of faults and anomalies. Therefore, the heterogeneous terminology used in the reviewed articles has been challenging for the extraction of each author's point of view on the anomaly detection approaches taxonomy. In this section, the authors' answer to RQ1.1 is presented in the form of a brief presentation of the categorizations found in the reviewed articles, followed by a detailed discussion of the works focused on each approach.

For instance, in [37], different dimensions are considered for the categorization of anomaly detection techniques for electric load consumption, such as the nature of the implemented artificial intelligence algorithm (i.e., supervised, unsupervised or semi-supervised), applications (i.e., energy saving, fault detection, theft attack detection, occupancy detection, at-home elderly monitoring), detection level (i.e., aggregated level, appliance level or spatio-temporal level) and computing platforms (i.e., edge computing, fog computing or cloud computing). However, there is no specific mention of the anomaly detection approach, which suggests that some authors consider it redundant with the dimension related to the algorithms and techniques. In [42], a statement is made about using terminology that refers to seven anomaly detection approaches, even though it does not match their definitions. In contrast, [41] only mentions four main categories, with many subcategories each. Furthermore, some articles suggest an initial dual classification, from which the remaining categories emerge, based on the differentiation between model-based and data-based approaches [43, 44], which will be discussed in Section 4.2.2. In [45], a distinct knowledge-based approach is presented. This method primarily depends on an in-depth understanding of battery mechanisms, along with knowledge and experience accumulated over time. It is particularly well-suited for complex, non-linear systems where mathematical modeling is not necessary.

Based on the findings regarding different anomaly detection approximations, common categories among different taxonomies have been identified, and additional categories have been incorporated to ensure a comprehensive representation of the identified approaches and algorithms. Consequently, from the review of the current state of the literature, five main categories have been established:

- Statistical approaches.
- Nearest neighbor-based approaches.
- Isolation-based approaches.
- Subspace-based approaches.
- Information-theoretic approaches.

4.2.1.1 Statistical approaches

The statistical approach is found to be notably present in the literature, particularly in publications focusing on large-scale systems [46, 47], emphasizing its

simplicity and efficiency for monitoring complex systems, as well as its robustness against unstable data quality and low resource deployment environments, which limit the full utilization of existing methods such as model-based and machine learning algorithms [47]. Other studies, such as Refs. [48, 49], also refer to the concept of robustness in applications related to power systems, thereby supporting the potential relationship between robustness-oriented approaches and statistical methods and techniques.

4.2.1.2 Nearest neighbor-based approaches

The nearest neighbor approach is frequently represented in the reviewed articles for different applications. In [50], it is presented as an outlier detection approach in a framework focused on short-term individual residential load forecasting [50]. In other works, it is innovatively combined with supervised learning techniques, primarily random forest (RF) [40] and SVM [51], or alternatively with algorithms from other fields of knowledge, such as the rain flow counting method used to assess the fatigue of a component under variable stress over time [52], with a focus on detecting collective anomalies in energy consumption [53]. Additionally, two publications [54, 55] refer to anomaly detection methodologies based on density-based methods, which were considered as a subcategory within the nearest neighbor in a longstanding publication [1].

4.2.1.3 Isolation-based approaches

Numerous articles refer to isolation-based methods, either as part of their anomaly detection framework or as reference algorithms for comparing the results of proposed approaches. In Refs. [50, 56], the isolation forest algorithm is employed as the primary technique for anomaly detection. Moreover, in Refs. [57, 58], isolation forest is utilized as a benchmark to evaluate the performance of the proposed methodology in anomaly detection for electricity consumption data. In particular, Gao et al. [59] propose an intelligent framework whose core technology is the isolation forest algorithm for ambient mode extraction for the smart grid. The use of such methods in high-impact articles suggests a growing interest in these algorithms within the energy sector, potentially indicating a degree of maturity, given that the approach has been in existence for over a decade [17].

4.2.1.4 Subspace-based approaches

As it was pointed out in Section 2, spectral anomaly detection approximations have been referred to in recent studies with the term *subspace*. Thus, whilst the term *spectral* is used in some of the reviewed articles, it is used when referring to spectral features or spectral clustering techniques. Conversely, subspace is mentioned in a number of articles, mainly referring to the concept of subspace, and only in one article [51] is a subspace detection method presented. Therefore, according to the reviewed articles, this approach does not seem to be considered in energy-related applications.

4.2.1.5 Information-theoretic approaches

The concept of information theory appears only once [44] in the reviewed articles, where mutual information theory and Gaussian copula entropy are applied

to examine the relationships between different condition monitoring variables and performance indicators of abnormal cases within a parameter identification framework for wind turbines. In this regard, a conclusion similar to the one presented for subspace-based approaches could be drawn.

Finally, approaches such as classification or clustering were not found to be presented solely anomaly detection approach; rather, these approaches were implemented in combination with other types, leading in some cases to detection and classification methodologies in applications related to wind turbines [44, 57] or power systems [60]. In fact, innovative combinations of up to five data-based methods have been proposed for the early detection of anomalous lithium-ion battery degradation [61]. These publications often refer to this approach as an ensemble, which will be discussed in Section 4.2.2 and which conforms to the majority group among the reviewed articles.

In summary, three of the five approaches identified (i.e., statistical, nearest neighbor and isolation) in the literature are present in three of the four most popular topics in the selected articles, suggesting that they are consolidated in the current state of the art regarding anomaly detection in the energy sector. In this sense, isolation-based methodologies may be emerging as the new benchmark with respect to the statistical approach, which has traditionally been considered the reference given its simplicity and potential for working with large volumes of data. In particular, the statistical approach is considered superior in long-term performance, arguing that artificial intelligence-based models struggle to cope with the variation in operating conditions that represent both normal and abnormal operating states due to system wear [62]. Interestingly, only a specific reference is made to the problem of collective anomaly detection, which is treated from the nearest neighbor perspective, which suggests a methodological gap.

Furthermore, it must be noted that the taxonomy of anomaly detection approaches is closely intertwined with the terminology referring to the types of problems to be solved (e.g., classification, regression), also depending on the nature of the data considered (e.g., unsupervised, supervised). Moreover, the fact that the presence of artificial intelligence is increasing in the field of anomaly detection cannot be ignored, and numerous studies on anomaly detection in other application areas place artificial intelligence on the same hierarchical level as approaches such as those identified in this study.

For illustrative purposes, the distribution of a number of references among the selected articles across the five identified categories regarding anomaly detection approaches, which focus on the four main topics covered by the reviewed studies, is presented in **Table 1**.

4.2.2 Algorithms and techniques

Anomaly detection approaches usually combine different algorithms and techniques in accordance with criteria such as real-time operation, availability of labeled data or prior knowledge of the system to which they are applied. Therefore, it is important to understand which algorithms and techniques were selected and the rationale that justifies their suitability in each case. In this section, the authors' answer to RQ1.2 is presented in the form of a brief discussion about the nature of the technique implemented and in the reviewed articles, followed by a discussion of some representative works on each approach.

Among the reviewed articles, a wide variety of techniques and algorithms have been identified, ranging from traditional model-based approaches to innovative

	Building electricity consumption data	**Heat distribution networks**	**Lithium-ion batteries**	**Renewable energy generation systems**
Statistical	[37]	—	[41]	[46, 63–65]
Nearest neighbor	[37]	—	[66, 67]	[51]
Isolation	[56]	[42]	—	[57]
Subspace	—	—	—	—
Information-theoretic	—	—	—	[44]

Table 1.
Distribution of anomaly detection topics across different approaches.

methods that combine data-driven techniques. The following classification encompasses the vast majority of techniques found in the reviewed works, specifically categorized into seven groups:

- Model-based methods.
- Statistical tests and analysis.
- Combined models, also known as *ensemble models*.
- Deep learning techniques.
- Data mining techniques.
- Innovative learning strategies.

4.2.2.1 Model-based methods

Model-based anomaly detection methods are considered among the traditional approaches for anomaly detection, together with manual analysis and thresholds [42]. Specifically, model-based anomaly detection approaches were used in only two of the reviewed articles [49, 60], which converge in their application to power systems state estimation. Besides state estimation methods, model-based anomaly detection methods include the parameter estimation method and the coordinated estimation method [41]. In [49], robust particle filters (PFs) are introduced, which are a type of sequential Monte Carlo (MC) algorithm used to estimate the state variables of dynamic systems, assuming that there can be errors or perturbations in the available observations [68]. In [60], an anomaly detection method is developed combining conventional weighted least squares (WLS) with extended Kalman filters (EKFs).

A distinct publication [55] integrates the theoretical model related to the power of a wind turbine in the outlier detection method, which uses the well-known density-based spatial clustering of applications with noise (DBSCAN) algorithm. As a result, a model-data hybrid-driven (MDHD) outlier detection method is presented for the wind turbine power curve (WTPC), which plays an essential role in many fields such as power forecasting and control [69, 70].

4.2.2.2 Statistical tests and analysis

On the topic of wind turbines, data-driven techniques have been used extensively in the reviewed articles for power curve modeling, condition monitoring and fault detection. **Figure 6** depicts the typical power curve shape for wind turbines.

Specifically, a number of publications that use statistical tests and analysis for wind turbines are found within the selected studies. For instance, Ohunakin et al. [65] apply the Kolmogorov-Smirnov test to evaluate turbine data for fault detection. Alternatively, in [63], a monitoring method based on stationarity is introduced, which relies on a sliding window approach and employs the Augmented Dickey-Fuller (ADF) test to examine the stationarity of data at each update step, where the resulting t-statistics are associated with fault identification. In [43], through a nonlinear autoregressive exogenous model (NARX), which is a statistical model for time series modeling, and previous knowledge fusion, the singular features of healthy gas turbines are revealed, and robust and sensitive anomaly detection is performed.

4.2.2.3 Combined models

The reviewed articles feature multiple data-driven techniques and algorithms, generally combined with each other to complement the strengths of different approaches. Thus, studies that employ a single technique are rare, such as [71], where a methodology is presented to integrate multiple data sources for fault diagnosis using a one-class support vector machine (OCSVM) classifier to assess normal behavior model error. Another example is [57], which exclusively uses the isolation forest technique for anomaly detection and removal. A comparative study of four anomaly detection methods (i.e., *iForest*, LOF, Gaussian mixture models (GMMs) and k-NN) for wind turbine power curve cleaning is presented in [72], highlighting Gaussian mixture models (GMMs) due to their favorable accuracy while maintaining wind variability. In contrast, Khan and Byun [73] suggest a new approach to detect anomalies in wind turbines using a combination of techniques, namely PCA, k-means clustering for labeling and a combination of classifier models in an ensemble scheme for outlier identification. In [74], an ensemble framework based on extreme gradient

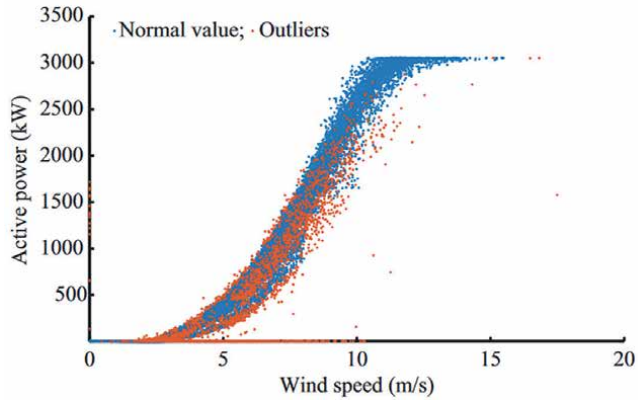

Figure 6.
Power curve modeling for wind turbine, final result of an outlier detection process [55].

boosting (*XGBoost*), integrating multiple machine learning and data mining techniques, was successfully developed for fault detection.

In lithium-ion battery-focused applications, different approaches to outlier and anomaly detection were also observed. In [66], two variants of the LOF algorithm were proposed and evaluated individually. Alternatively, in [61], the LOF algorithm was combined with four other methods of different natures and used to develop an ensemble-based algorithm to robustly identify anomalous samples as early as possible. Specifically, these methods include regression models with prediction bounds, SVM, Mahalanobis distance (MD) and sequential probability ratio test (SPRT).

The combination of models in ensemble schemes also enriches other application domains, such as building energy consumption monitoring. In [75], the use of two complementary semi-supervised machine learning applications, based on classification and regression tree (CART) and multi-layer perceptron (MLP), is proposed to achieve highly interpretable and accurate anomaly detection. Lei et al. [76] introduce a dynamic anomaly detection algorithm tailored for building energy consumption data, capable of identifying both point anomalies and collective anomalies. This approach integrates supervised clustering with an unsupervised method, where particle swarm optimization (PSO) is employed to fine-tune the parameters of the unsupervised clustering algorithm.

4.2.2.4 Deep learning techniques

Some publications that investigate anomalies in electricity consumption in buildings or that consider the resilience of various infrastructures, such as advanced metering infrastructure (AMI) and heating, ventilation and air conditioning (HVAC) systems, focus on implementing deep learning techniques. In [77], an asymmetric hybrid encoder-decoder (AHED) architecture is presented for anomaly detection, aimed at accurately predicting and identifying both point and collective anomalies in the context of building energy consumption. This framework combines supervised and unsupervised learning techniques and employs a sophisticated encoder-decoder structure to enhance the precision of energy usage forecasting. In contrast, regarding the resilience of infrastructures, Elnour et al. [56] develop and validate a semi-supervised, data-driven attack detection strategy using an isolation forest, combined with deep learning techniques for temporal feature extraction, specifically a 1D CNN-based encoder. In [78], profiles obtained from advanced metering infrastructure (AMI) meters are used to create 2D images through a continuous wavelet transform. Subsequently, several deep learning models are sequentially applied for feature extraction and the detection of false data injection attacks (FDIA). Similarly, in [45], a sparse autoencoder is used to extract the characteristic parameters of battery faults from the reconstructed high-frequency part of the original voltage signal.

Long short-term memory (LSTM), a deep learning algorithm, is increasingly recognized in the literature as a powerful technique for anomaly and fault detection in wind turbine applications. In [44], an approach combining LSTM and AE neural networks is proposed to evaluate sequential condition monitoring data from wind turbines. This method builds a performance assessment model using LSTM units and AE networks to compute performance indices, which are used to quantify the degree of performance anomalies. From this, key condition monitoring parameters are identified by analyzing their relationships with abnormal performance instances. Identifying these critical parameters is essential not only for detecting potential faults in wind turbines but also for optimizing the use of limited fault data to identify issues

across different wind turbine generators (WTGs), especially when data distributions differ between units. Additionally, as discussed in [79], the scarcity of fault data in real-world wind turbine operations has led to growing interest in transfer learning (TL) for condition monitoring and fault diagnosis. In this regard, Zhu et al. [79] present a hybrid method that integrates LSTM, fuzzy synthesis and feature-based TL.

4.2.2.5 Data mining techniques

Regarding data mining techniques and methodologies, some of the reviewed articles consider bootstrapping for anomaly detection, which is a data resampling method for estimating the distribution of a statistic. Specifically, in [80], it is a central concept, as it is used to improve a traditional threshold-based outlier detection method, which is local correlation integral (LOCI), as it focuses on searching for an improvement on the thresholds used in the classification of the LOCI method by analyzing its distribution. This approach helps eliminate subjectivity in threshold selection by data analysts or maintenance personnel, with the ultimate goal of improving energy efficiency in building operations.

4.2.2.6 Innovative learning strategies

It has been observed that among the reviewed works, some particularly focus on innovative learning strategies. For instance, Choubey et al. [81] address scenarios characterized by limited and low-quality data, aiming to reduce dependence on large, balanced and labeled datasets with complex patterns. By incorporating advanced feature extraction techniques, the authors propose a contrastive learning model that improves the performance, reliability and scalability of electricity load anomaly detection. This approach enhances cost-effectiveness and adaptability across a wider range of real-world applications.

Other publications consider anomaly detection schemes that take into account the evolution of the systems being monitored. In [82], the challenge of detecting evolving electricity theft behaviors in modern power systems is addressed through a combination of active learning and incremental learning. The proposed model integrates active learning with incremental support vector data description, using an adaptive mechanism to identify candidate support vectors and incrementally update the existing model. This strategy effectively balances computational efficiency and detection accuracy, accommodating the evolving nature of electricity theft.

Similarly, [83] focuses on the nonstationary behavior of lithium-ion battery cells during charging and discharging processes, which complicates anomaly detection. To address this, the authors introduce a condition-driven mode partition strategy that identifies multiple operational modes within the nonstationary data, enabling more effective anomaly detection under varying operational conditions.

Summarizing, among the five identified categories regarding anomaly detection techniques (i.e., model-based, statistical, ensemble, DL-based and data mining), only the combination of different techniques and algorithms (i.e., combined models or ensemble) is present in the four most represented topics in the selected articles, which supports that anomaly detection schemes based on an ensemble of models are popular across diverse energy applications. In fact, despite there being longstanding publications that refer to ensemble learning, the open and flexible nature of it, which allows for the combination of innovative algorithms and techniques as well as for innovative combination schemes, promotes its continued relevance.

Now, regarding specific types of algorithms and techniques, DL techniques are found to be employed in three of the four main topics, only missing in the heat distribution networks. This is coherent with the increasing use of DL over ML in many applications, due to the ability to detect hidden patterns in complex datasets without the need for previous feature selection. Therefore, DL becomes suitable for anomaly detection applications, such as cyber-attack detection and fault detection, which is supported by the publications referring to data-driven cyber-attacks toward electric buildings, as well as by the studies that focus on condition monitoring on wind turbines and electric vehicle batteries. Also, DL techniques are specifically chosen for the treatment of spatio-temporal information for anomaly detection in [58], where multi-scale graphs are used to learn spatial features, and convolutional networks are proposed for learning the temporal features. However, a different paper among the selected studies [84] proposes an alternative method for spatio-temporal analysis based on inverse distance weighting (IDW), questioning the need to use complex machine learning or deep learning algorithms to tackle this kind of problem, including considerations for data-constrained scenarios in which meteorological data such as irradiance and temperature is unavailable.

There are only two references in the selected literature that propose model-based anomaly detection, which converge on power state estimation. This suggests that data-driven approaches are more effective for anomaly detection applications and do not require modeling expertise, despite sometimes compromising the explainability of the results.

Finally, within the reviewed works, some focus on more singular techniques, as well as innovative learning strategies. On one hand, data mining techniques are proposed in only one study, suggesting that, despite bringing an interesting big data perspective, which can work as an alternative to traditional methods, other data-driven algorithms such as ML or DL are preferred by the researchers. On the other hand, innovative learning strategies are focusing on specific scenarios in which training anomaly detection models is particularly complex, such as low data quality and availability, dynamic environments and systems with multiple operation conditions, all of which are interesting in the energy application domain.

For illustrative purposes, the distribution of a number of references among the selected articles across the five identified categories regarding anomaly detection techniques, which focus on the four main topics covered by the reviewed studies, is presented in **Table 2**.

	Building electricity consumption data	**Heat distribution networks**	**Lithium-ion batteries**	**Renewable energy generation systems**
Model-based	—	—	—	[55]
Statistical	[37]	—	[41]	[46, 63–65]
Combined models	[75, 76]	[85]	[61, 66]	[71, 73, 74]
Deep learning	[58, 78]	—	[45]	[44, 79]
Data mining	[80]	—	—	—

Table 2.
Distribution of anomaly detection topics across different types of algorithms and techniques.

4.2.3 Layer of application

Anomaly detection measures in energy applications are deployed using various techniques and algorithms, which can be implemented at different scopes that can be classified into hierarchical levels (i.e., component, device, system, process and plant). Therefore, it is relevant to analyze how different approaches and techniques for anomaly detection relate to these levels. In this section, the authors present their answer to RQ1.3 in the form of a brief discussion, moving from the lower hierarchical level to the higher level.

A particular article [86] stands out as the most specific case by focusing directly on a single system component—high-frequency-link power conversion systems. These systems benefit from recent advances in data-driven modulation approaches, which aim to automate design processes. However, the performance of data-driven models can degrade significantly in the presence of outliers or when training data is insufficient. To mitigate these challenges, the study introduces an artificial intelligence-based methodology for outlier detection tailored to this application.

In the field of electricity consumption measurement, a few works were found to take a slightly broader approach than the previously presented. Instead of focusing on the integrated circuit within a device, they analyze the signal that the device is capable of measuring. For example, Li et al. [87] propose a domain knowledge-based and topology-aware anomaly detection algorithm that uses sensor data from a dynamic grid. This method integrates time series data of both measurements and topological changes, employing graph distances informed by domain knowledge to estimate reliable distributions of measurements at each time step. From a different perspective, in [78], a semi-supervised scheme is presented, which utilizes the ratio profile generated from the readings of the observer meter and the user's smart meter as the input, which is processed in an innovative manner in order to reduce false positives (FPs) for energy theft. Among the reviewed articles, one study specifically considers increasing granularity from the perspective of consumption disaggregation. Lastly, Castangia et al. [88] take a fine-grained approach to anomaly detection through energy consumption disaggregation. It presents a method for identifying electrical faults in household appliances by analyzing their unique power signatures, highlighting the potential of detailed consumption data to support targeted fault detection in residential settings.

Regarding anomaly detection approaches for more complex devices, a considerable number of works within the reviewed articles integrate outlier or anomaly detection in the management of batteries, which require specific considerations due to their chemical nature, or wind turbines, which are large-scale electromechanical devices. As previously discussed, anomaly detection and classification for lithium-ion batteries, usually performed by the BMS, are key to ensuring safe and reliable operation, a fact supported by the battery-focused publications among the reviewed articles [41, 45, 47, 66, 83]. Similarly, as previously mentioned, the characterization and monitoring of wind turbines have garnered significant interest among the reviewed articles, enabling more efficient utilization of renewable energy.

However, a notable area of interest in anomaly detection focuses on processes, plants or systems in which early identification of anomalies in large-scale infrastructures helps prevent critical consequences, rather than detecting the failure of a single component or device. Among the reviewed articles, a number of references directed at the system level have been identified, specifically regarding the management of HVAC systems, both from the perspective of detecting specific faults in air handling units

(AHUs) and rooftop units (RTUs) [89], as well as from the standpoint of cybersecurity [56] and the management of substations within a district heating network [42].

In [54], the focus is placed on the analysis of blast furnace gas (BFG) and Linz-Donawitz converter gas (LDG), which are typical by-product gases generated during the iron- and steel-making processes [90]. The generation and consumption flows, as well as the variations in the gas tank levels, are intrinsically related due to the underlying manufacturing processes. For instance, in the BFG system—characterized by a continuous production process—the complexity of the blowing-down and re-blowing mechanisms causes severe fluctuations in the gas generation flow. These fluctuations often lead to inconsistencies between the sensor readings and the actual process conditions, resulting in a significant number of anomalous data points. On the other hand, the LDG system follows a batch process [91], with intermittent off-recycling stages between nitrogen blowing and slag splashing cycles designed to protect the converters [92]. Due to the complexity of molten steel composition and the fluctuations in oxygen flow, abnormal data are commonly found in the sensor measurements related to LDG generation flow.

Lastly, two publications among the reviewed articles have been identified that consider a plant-level scope, meaning they are capable of detecting anomalies in large-scale infrastructures. In [64], an integrated diagnostic pipeline is proposed for PV systems, combining various innovative routines to differentiate between common failures and performance degradation modes—such as zero or reduced power output, system degradation, soiling and snow-related losses. This approach relies on a single performance metric and is intended for both batch analysis of large PV fleets and real-time monitoring, provided that the technical specifications (e.g., system characteristics and meteorological data) are available. Similarly, Wang et al. [47] introduce a comprehensive data-driven assessment framework designed for multitask management within cloud-based battery management systems. This approach is aimed at enhancing the overall performance and scalability of such systems, particularly in the context of lithium-ion batteries used in electric vehicles (EVs), where integrating multiple tasks efficiently is a growing area of interest.

Summarizing, most of the selected studies focus on device, system and process levels, proposing a range of anomaly detection approximations and techniques that have been discussed in previous sections. Within them, HVAC machinery, EV batteries and wind turbines are the most popular devices and systems found in the selected studies. Contrarily, few studies address either component level (e.g., power converters) or large-scale architectures (i.e., plants), suggesting that both ends of the spectrum are more niche knowledge fields and publications are scarce.

Regarding approximations and techniques, statistical tests and analysis span across the hierarchical levels and are specifically chosen for the large-scale scenarios. Differently, DL-based and model-based approaches focus on more intermediate layers (i.e., device, system and process).

4.2.4 Datasets

The rise of data-driven methods reinforces the importance of the availability and quality of representative data for the component, device, system, process or plant to be monitored from the perspective of fault or anomaly detection. Therefore, it is important to investigate various aspects of datasets used for the validation of algorithms and techniques in anomaly detection applications. In this section, the authors present their answer to RQ1.4 in the form of a brief discussion.

Datasets play a crucial role throughout the development process, whether they serve as proof of concept for visualization, manual examination and statistical analysis or as training input for machine learning techniques like regression, clustering and classification [42]. In a notable number of the reviewed articles, real datasets are used for validation purposes, which were usually presented together with the main characteristics of the component, device or system from which they originate. For example, in publications referring to renewable energy generation infrastructures, details such as the peak power generation capacity of a PV plant or a wind turbine are typically provided. In particular, these studies often specify the location of the infrastructure, whereas only in some cases, such as Refs. [44, 74], are details of the physical characteristics of the infrastructure included. Specifically, due to the emphasis on power curve modeling in studies focused on wind turbines, some of these articles also include details regarding the configuration of their transmission systems [63]. Special consideration is given to articles that used open datasets for residential load forecasting [50], electricity theft detection in smart grids [78, 82] and wind turbine monitoring [63].

In some cases, there is a notable interest in validating the proposed algorithms and methodologies using both real and simulated data as an effort to demonstrate their applicability. For instance, in [66], both simulation data from an air-cooled lithium-ion battery energy storage system and experimental data from a water-cooled lithium-ion battery energy storage system are considered. Similarly, in [89], the dataset consists of both simulated (i.e., modeled) and experimental (i.e., physical) data from test facilities. Additionally, some references consider standardized IEEE test cases, such as the IEEE 16-generator 5-area system, which is examined in [59]; the IEEE 39-bus New England system, which is analyzed in [49]; the IEEE 14-bus test system, considered in [60]; and the IEEE 33-bus power distribution system, studied in [48]. In contrast, two articles focused on power system state estimation [49, 60] rely exclusively on simulation data. However, it is important to highlight that these simulation datasets correspond to standardized IEEE test cases.

A few works among the reviewed articles focus only on experimental data obtained from an on-premises experimental testbed. For instance, in [67], an experimental platform is used for battery testing in different temperature ranges, which can be controlled by a temperature chamber. Similarly, a prototype platform was used in [86] for experimental verification on dual active bridge power converter monitoring.

In summary, the proliferation of data-driven approaches, techniques and algorithms is highlighting the importance of open, rich and representative datasets for research and development of anomaly detection in the energy applications domain. Although high-impact researchers support their studies with a high level of detail when describing the datasets and infrastructures from which data has been acquired, there is still a notable amount of research that uses longstanding datasets, which could question the applicability of the proposed developments.

4.2.5 Evaluation metrics

Evaluation metrics are key to assessing the performance of existing solutions, as well as for benchmarking against novel developments. Within the selected studies, those that utilize labeled datasets often utilize a range of evaluation metrics or apply the same metrics in varying manners. Specifically, some metrics appear to be used by most of the authors, specifically for classification and forecasting problems.

Several studies use classic metrics related to binary classification problems, which are fundamental in anomaly detection applications, in which normal data points are assigned to the negative class, and anomalous instances are classified as the positive class.

From this perspective, different metrics are calculated from the number of occurrences of correct predictions on the positive samples, true positive (TP), and negative samples, true negative (TN), and from the number of occurrences of incorrect predictions on both conditions, false positive (FP) and false negative (FN), respectively.

Precision, or positive predictive value (PPV), is computed by dividing TP occurrences by the total number of samples classified as positive by the model, which describes the sterility of the detected faults and whether there are FPs present. Precision is one of the most popular metrics within the reviewed studies [56, 73, 77, 93].

Alternatively, Recall, Sensitivity or true positive rate (TPR), is computed by dividing TP by all of the positive samples that were used for validation, which describes the rate of existing faults detected by the classifier. This metric is considered together with Precision a number of times [73, 93], while other studies [60] opt to combine them into a single metric, which is the F1-score, computed as the harmonic mean of Precision and Recall, as there is a severe class imbalance in the considered dataset. TPR can also be investigated against the probability of false alarm or false positive rate (FPR), which leads to the receiver operating characteristics (ROC) curve, describing how the ratio of TP and FP shifts across all thresholds. Thus, the area under the ROC curve (ROC-AUC) is sometimes used as a summarized indicator of model quality [42].

A particular paper [82] on studying electricity theft utilizes Specificity, or true negative rate (TNR), which is the negative equivalent of Sensitivity, considering that TN indicates the number of samples of electricity theft customers with correct detection results. Therefore, TNR quantifies the proportion of correctly classified theft samples among all actual theft cases.

Moreover, a few references use Accuracy [64, 89], which is computed by dividing the sum of the two elements on the diagonal of the confusion matrix (TP and TN) by the total number of positive and negative samples, measuring the percentage of samples correctly classified. A specific study [94] considers balanced accuracy (BA), which provides a balanced measure of classification performance by considering both Sensitivity and Specificity, which is used in the presence of imbalanced data. Finally, studies that include metrics for forecasting performance [50, 95] propose the mean absolute percentage error (MAPE).

In summary, a range of metrics is identified among all the reviewed studies, but many of them underscore the importance of the development of unified metrics that enable trustworthy and comparable evaluation of already existing solutions.

4.3 Open issues and research agenda

Drawing from the results of research question RQ1, additional conclusions are presented that highlight application areas and methodological approaches that are rarely addressed or considered in the reviewed studies.

First, it is found that two relevant topics are missing within the reviewed papers, suggesting that they are under-represented trends in the sample of studies, which has resulted from the selection process. On one hand, there are no papers specifically addressing anomaly detection in critical infrastructures and critical entities, which provide essential services to society in sectors such as energy, transport, finance or health, among others. In that context, the interdependence of many sectors on the

energy system (i.e., power generation, transportation and distribution), such as the digital infrastructure for communications, underscores the need to develop specific anomaly detection applications for critical infrastructures of the energy sector. Notably, given their critical nature, it must be noted that standards and regulations regarding critical infrastructures are more mature when compared to other application domains, which could discourage researchers who are not willing to take the time to understand the legal and regulatory framework. On the other hand, the increased penetration of renewable energy, along with the proliferation of electric vehicles, has promoted the increase in charging points throughout the country, offering more charging options to EV users and also enabling flexibility solutions that integrate energy exchanges between the EV and the distribution grid, sometimes referred to as vehicle-to-grid (V2G). However, no publications within the selected studies address anomaly detection for EV chargers or EV charging stations.

Second, a number of methodological gaps are identified when carefully studying the selected works. For instance, federated learning (FL), which is an emerging ML paradigm that enables a large number of actors to perform an on-device training of a single ML model without sharing raw data, is only mentioned once in the 52 papers. However, FL is believed to be very powerful as infrastructures across all sectors become more and more distributed, which also includes the energy sector. In parallel, the usage of generative IA is scarce, mainly in the form of GANs and GAN-based algorithms, and focused on extending the available dataset, as well as generating imbalanced data.

Furthermore, DL-related methodological gaps have been found. On one hand, from a methodological point of view, it is not clear how to tackle the resource consumption, explainability and applicability of DL algorithms in anomaly detection applications. On the other hand, algorithms derived from DL, such as deep reinforcement learning (DRL), are emerging in other engineering applications and sectors, but there are very few references to them among the selected articles. In fact, the suitability of DRL is explicitly highlighted only in the context of detecting more complex anomalies, which typically involve high-dimensional data, such as consumption patterns and environmental conditions, as well as challenges like uncertain agent observations and sparse reward signals for anomaly identification [37]. In all other cases, DRL is only referred to in future works and future research directions.

Now, the current open issues of anomaly detection research are pointed out in order to answer RQ2. For each open issue, some research directions are proposed to address it.

4.3.1 Real-time operation and online deployment

Real-time operation is a key aspect of anomaly detection systems in energy applications. Specifically, it is critical in distributed environments such as the energy metering infrastructure or distributed renewable energy resources, which are equipped with resource-constrained edge devices that are limited in terms of computational resources and storage. Reducing training expenses of anomaly detection models, increasing detection efficiency and allowing real-time data analysis are believed to transform anomaly detection systems into more dependable and scalable solutions for many real-world anomaly detection situations [81].

Cloud-edge collaborative frameworks, such as those proposed by Li et al. [41], in which deep learning models can be deployed on both edge devices and cloud servers, enable efficient data processing and analysis. Thus, edge computing and model

compression techniques can address the limitations of computational resources and bandwidth in vehicular environments, while cloud computing provides a robust platform for more complex battery anomaly detection tasks, including training deep learning models, long-term data storage and historical data analysis.

4.3.2 Explainability and interpretability

In recent years, the enduring challenge of evaluating and quantifying machine learning explainability has garnered some attention [96]. Nonetheless, many of the existing approaches are tailored primarily to classification or clustering tasks, making their adaptation to anomaly detection scenarios particularly challenging [97]. As a result, the interpretability of anomaly detection techniques has gained growing significance.

Therefore, explainable anomaly detection (XAD) refers to the process of deriving meaningful information from an anomaly detection model regarding patterns present in the data or acquired by the model. This information is deemed relevant when it offers valuable insight into the anomaly detection problem being examined by the end-user [97]. However, authors in [98] argue that most existing outlier detection methods typically fail to provide explanations for why certain instances are classified as outliers, that is, they do not clearly identify the specific characteristics that make those instances stand out.

Explainability is especially limited in deep learning-based anomaly detection approaches. Although these methods often deliver strong performance, the black-box nature of deep learning models poses a challenge for practical implementation [99], as the lack of transparency can reduce operator trust and hinder real-world deployment [94]. This issue is particularly relevant in the context of energy consumption anomaly detection, where understanding the reasons behind detected anomalies is essential. Therefore, developing deep learning-based methods that can explain why a particular power consumption event or observation is considered abnormal can help experts concentrate on the most critical anomalies and enhance their confidence in the implemented solutions [100, 101].

4.3.3 Learning strategies

Alternative and innovative learning strategies emerge in recent works in order to enable models to consider the evolution of different systems and the characteristics of anomalous data points, as well as to reduce the dependence on large, balanced, labeled datasets containing intricate patterns.

For instance, several recent studies have explored the use of TL strategies in condition monitoring and fault diagnosis, motivated by the limited availability of fault data during the actual operation of wind turbines [79].

Alternatively, by employing advanced feature extraction methods, contrastive learning models improve the effectiveness, robustness and scalability of electricity load anomaly detection, thereby increasing cost-efficiency and making them more suitable for a wide range of real-world applications [81].

In the context of energy theft detection, current approaches often struggle to effectively learn and adapt to the continuously changing and complex nature of theft behaviors. Moreover, they frequently fall short in meeting the real-time processing demands required for analyzing such behaviors. Research on incremental learning tailored to theft detection remains relatively limited [102]. Similarly, as discussed in

Section 4.2.2, specific methodologies need to be developed to cope with the variation in operating conditions that represent both normal and abnormal operating states due to turbine wear [62].

In general, future research needs to provide anomaly detection algorithms for energy applications with adaptive incremental learning capabilities to enhance the resilience and applicability of anomaly detection models.

4.3.4 Data availability and reproducibility

Although significant progress has been made in developing anomaly detection techniques for energy-related applications, several factors have been identified that hinder reproducibility and, consequently, the fair and consistent experimental comparison of these algorithms [37]. One of the main challenges lies in evaluating the generalizability of anomaly detection approaches, as most frameworks are typically tested on a single dataset, often supplied by a partnering utility company. Such datasets are usually unlabeled and not publicly accessible, as is the case for the majority of studies reviewed, as mentioned in Section 4.2.4. This lack of accessibility complicates method comparison, increases the risk of error, and ultimately slows down advancement in the field [42].

To address this challenge, there is a pressing need to make open-source anomaly detection toolkits available, which should include challenging energy-related datasets alongside existing anomaly detection algorithms. This would enable fair, straightforward and reproducible comparisons of different methods [37]. Additionally, experimental data can serve as a valuable alternative to simulation data for data-driven modeling. The potential of combining simulation data with experimental results in hybrid approaches should also be explored. Furthermore, advanced artificial intelligence techniques such as attention neural networks, physics-informed neural networks and multitask learning can be integrated to enhance the practicality and effectiveness of data-driven models [86].

As highlighted in Section 4.2.2, numerous reviewed articles integrate these techniques, particularly LSTM, AE and CNN. The strength of deep learning lies in its ability to handle large-scale datasets and automatically learn discriminative features from the data, removing the need for manual feature engineering by domain experts. However, despite these advantages, deep neural networks (DNNs) face challenges during training that make them vulnerable [103]. For example, training can take several hours or even days. Additionally, deep models commonly suffer from overfitting and require a large volume of samples to train effectively [104], often resulting in poorer performance compared to shallow machine learning methods when training data is limited.

The use of edge control devices equipped with advanced microcontrollers featuring integrated machine learning accelerators enables inference, and potentially training, to be carried out directly on small, resource-limited, low-power devices rather than relying on large computing systems (such as desktops or workstations) or cloud platforms. Consequently, deep learning models must be compressed to fit the limited computing power, storage capacity and bandwidth of these devices, all while preserving their core functionality and accuracy [37].

5. Conclusions and future work

In this study, the results of a systematic literature review on anomaly detection in energy applications have been presented. The results aim to shed some light on a

long-established research area, which has been developed within diverse research areas. More specifically, this work focused on answering two general research questions and the corresponding sub-questions, which are summarized as follows:

The main anomaly detection applications in the energy sector appear to be energy distribution networks, that is, heat and electricity, as well as renewable energy generation, mainly PV and wind energy, and energy storage systems.

Regarding the main anomaly detection methodologies identified in previous reviews and surveys, some of them show very low or no presence among the selected articles, which underpins the idea that new taxonomic approaches are needed to improve terminological consistency in the field.

Regarding the techniques and algorithms implemented in the selected articles, a broad range of approaches has been observed, primarily model-based and data-driven. Notably, there is a strong presence of artificial intelligence-based models, especially DL, as an emerging and powerful tool for anomaly detection, although some knowledge and methodological gaps arise related to it. Also, a significant use of statistical techniques and particularly isolation-based algorithms, such as isolation forest and its derivatives, has been identified. In some cases, algorithms have been deployed on cloud infrastructures, monitoring large-scale systems, while in others, models are implemented at the edge level, prioritizing decentralized and real-time anomaly detection.

Under-represented trends have been investigated and discussed, as well as open issues, not only related to edge-cloud collaboration and online model updating but also encompassing adaptive and incremental learning strategies, as well as the explainability and interpretability of models, particularly in the case of deep learning-based approaches. Furthermore, its improvement will encompass research reproducibility and reduce the likelihood of the same methodology being developed simultaneously and unknowingly by multiple researchers.

From a continuation perspective, anomaly detection is considered to have a cross-cutting impact across various application domains. While the theoretical foundation provided by the algorithms remains consistent, the challenges encountered and the methodological or procedural solutions required vary by sector. In this regard, the authors aim to extend this line of research by applying anomaly detection analysis to fields currently undergoing significant digital advancements, particularly cybersecurity in critical infrastructure, where the physical and cyber dimensions must be addressed within a unified framework.

Conflict of interest

The authors declare no conflict of interest.

Nomenclature

RQ	research question
SLR	systematic literature review
MNR	maximum normed residual
KDE	kernel density estimation
kNN	k-nearest neighbor
LOF	local outlier factor

COF	connectivity-based outlier factor
SOD	subspaces outlier degrees
NN	neural network
SVM	support vector machine
SOMs	self-organizing maps
EM	expectation maximization
DL	deep learning
CNN	convolutional neural network
RNN	recurrent neural network
AE	auto-encoder
GAN	generative adversarial network
ML	machine learning
PCA	principal component analysis
IC	inclusion criteria
EC	exclusion criteria
SJR	SCImago journal rank
DHS	district heating substations
PV	photovoltaic
EV	electric vehicle
BMS	battery management system
RF	random forest
PF	particle filter
MC	Monte Carlo
WLS	weighted least squares
EKF	extended Kalman filter
DBSCAN	density-based spatial clustering of applications with noise
MHDH	model-data hybrid-driven
WTPC	wind turbine power curve
ADF	Augmented Dickey-Fuller
NARX	nonlinear autoregressive exogenous model
OCSVM	one-class support vector machine
GMM	Gaussian mixture model
MD	Mahalanobis distance
SPRT	sequential probability ratio test
CART	classification and regression tree
MLP	multi-layer perceptron
PSO	particle swarm optimization
HVAC	heating, ventilation and air conditioning
AHED	asymmetric hybrid encoder-decoder
AMI	advanced metering infrastructure
FDIA	false data injection attacks
LSTM	long short-term memory
WTG	wind turbine generator
TL	transfer learning
LOCI	local correlation integral
IDW	inverse distance weighting
AHU	air handling unit
RTU	rooftop unit
BFG	blast furnace gas
LDG	Linz-Donawitz converter gas

TP	true positive
TN	true negative
FP	false positive
FN	false negative
PPV	positive predictive value
TPR	true positive rate
FPR	false positive rate
ROC	receiver operating characteristics
AUC	area under the curve
TNR	true negative rate
BA	balanced accuracy
MAPE	mean absolute percentage error
V2G	vehicle to grid
DNN	deep neural network
XAD	explainable anomaly detection
DRL	deep reinforcement learning

Author details

Joan Valls Pérez[1*], Mayra Ramírez Chávez[2], Miguel Delgado-Prieto[2] and Luis Romeral Martínez[1]

1 Electronics Engineering Department, MCIA Research Centre, Universitat Politècnica de Catalunya, UPC, Terrassa, Spain

2 Automatic Control Department, MCIA Research Centre, Universitat Politècnica de Catalunya, UPC, Terrassa, Spain

*Address all correspondence to: joan.valls.perez@upc.edu

IntechOpen

References

[1] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM Computing Surveys. 2009;**41**(3):1-58

[2] Grubbs FE. Procedures for detecting outlying observations in samples. Technometrics. 1969;**11**(1):1-21

[3] Hawkins DM. Identification of Outliers. Dordrecht: Springer Netherlands; 1980. Available from: http://link.springer.com/10.1007/978-94-015-3994-4

[4] Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: Identifying density-based local outliers. SIGMOD Record. 2000;**29**(2):93-104

[5] Hu T, Sung SY. Detecting pattern-based outliers. Pattern Recognition Letters. 2003;**24**(16):3059-3068

[6] Nassif AB, Talib MA, Nasir Q, Dakalbab FM. Machine learning for anomaly detection: A systematic review. IEEE Access. 2021;**9**:78658-78700

[7] Ariyaluran Habeeb RA, Nasaruddin F, Gani A, Targio Hashem IA, Ahmed E, Imran M. Real-time big data processing for anomaly detection: A survey. International Journal of Information Management. 2019;**45**:289-307

[8] DeMedeiros K, Hendawi A, Alvarez M. A survey of AI-based anomaly detection in IoT and sensor networks. Sensors. 2023;**23**(3):1352

[9] Wu Y, Dai HN, Tang H. Graph neural networks for anomaly detection in industrial internet of things. IEEE Internet of Things Journal. 2022;**9**(12):9214-9231

[10] Samariya D, Thakkar A. A comprehensive survey of anomaly detection algorithms. Annals of Data Science. 2021;**10**:829-850. Available from: https://link.springer.com/10.1007/s40745-021-00362-9

[11] Pang G, Shen C, Cao L, Hengel AVD. Deep learning for anomaly detection: A review. ACM Computing Surveys. 2021;**54**(2):1-36. DOI: 10.1145/3439950

[12] Xia X, Pan X, Li N, He X, Ma L, Zhang X, et al. GAN-based anomaly detection: A review. Neurocomputing. 2022;**493**:497-535

[13] Che TC, Duan HF, Lee PJ. Transient wave-based methods for anomaly detection in fluid pipes: A review. Mechanical Systems and Signal Processing. 2021;**160**:107874

[14] Ahmed M, Naser Mahmood A, Hu J. A survey of network anomaly detection techniques. Journal of Network and Computer Applications. 2016;**60**:19-31

[15] Ahmed M, Mahmood AN, Islam MR. A survey of anomaly detection techniques in financial domain. Future Generation Computer Systems. 2016;**55**:278-288

[16] Cook AA, Misirli G, Fan Z. Anomaly detection for IoT time-series data: A survey. IEEE Internet of Things Journal. 2020;**7**(7):6481-6494

[17] Liu FT, Ting KM, Zhou ZH. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data. 2012;**6**(1):1-39

[18] Ghoting A, Otey ME, Parthasarathy S. LOADED: Link-based outlier and anomaly detection in evolving data sets. In: Fourth IEEE International Conference on Data Mining (ICDM'04). Brighton, UK: IEEE;

2004. pp. 387-390. Available from: http://ieeexplore.ieee.org/document/1410317/

[19] He Z, Deng S, Xu X. A Unified Subspace Outlier Ensemble Framework for Outlier Detection. In: Fan W, Wu Z, Yang J, Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, et al., editors. Advances in Web-Age Information Management, Lecture Notes in Computer Science. Vol. 3739. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. pp. 632-637. Available from: http://link.springer.com/10.1007/11563952_56

[20] Abe N, Zadrozny B, Langford J. Outlier detection by active learning. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia PA USA: ACM; 2006. pp. 504-509. Available from: https://dl.acm.org/doi/10.1145/1150402.1150459

[21] Chandola V. Outlier Detection: A Survey. Available from: https://www.researchgate.net/profile/Vipin-Kumar-54/publication/242403027_Outlier_Detection_A_Survey/links/0deec52d83600bb86c000000/Outlier-Detection-A-Survey.pdf

[22] Bellman R. Dynamic programming. Science. 1966;**153**(3731):34-37

[23] Zhang L, Lin J, Karim R. Sliding window-based fault detection from high-dimensional data streams. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2016;**47**:1-15

[24] Thudumu S, Branch P, Jin J, Singh J. A comprehensive survey of anomaly detection techniques for high dimensional big data. Journal of Big Data. 2020;**7**(1):42

[25] Xu X, Liu H, Yao M. Recent progress of anomaly detection. Complexity. 2019;**2019**(1):2686378

[26] Cao Y, Xiang H, Zhang H, Zhu Y, Ting KM. Anomaly detection based on isolation mechanisms: A survey. arXiv. 2024. Available from: https://arxiv.org/abs/2403.10802

[27] Toshniwal A, Mahesh K, R. J. Overview of anomaly detection techniques in machine learning. In: 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). Palladam, India: IEEE; 2020. pp. 808-815. Available from: https://ieeexplore.ieee.org/document/9243329/

[28] Samara MA, Bennis I, Abouaissa A, Lorenz P. A survey of outlier detection techniques in IoT: Review and classification. Journal of Sensor and Actuator Networks. 2022;**11**(1):4

[29] Marchioni A, Enttsel A, Mangia M, Rovatti R, Setti G. Anomaly detection based on compressed data: An information theoretic characterization. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2024;**54**(1):23-38

[30] Moustafa N, Hu J, Slay J. A holistic review of network anomaly detection systems: A comprehensive survey. Journal of Network and Computer Applications. 2019;**128**:33-55

[31] De Almeida Biolchini JC, Mian PG, Natali ACC, Conte TU, Travassos GH. Scientific research ontology to support systematic review in software engineering. Advanced Engineering Informatics. 2007;**21**(2):133-151

[32] Kitchenham B, Charters S. Guidelines for Performing Systematic Literature Reviews in Software Engineering. EBSE Technical Report EBSE-2007-01. United Kingdom: Keele University

[33] Kitchenham BA, Budgen D, Pearl BO. Using mapping studies as the basis for further research—A participant-observer

case study. Information and Software Technology. 2011;**53**(6):638-651

[34] SCImago Journal Rank (SJR) [Internet]. Available from: https://www.scimagojr.com/

[35] ACM DL [Internet]. Available from: https://dl.acm.org/

[36] Scopus [Internet]. Available from: https://www.elsevier.com/products/scopus/search

[37] Himeur Y, Ghanem K, Alsalemi A, Bensaali F, Amira A. Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. Applied Energy. 2021;**287**:116601. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100478413&doi=10.1016%2fj.apenergy.2021.116601&partnerID=40&md5=cfcaf88aa168a4b327e69d1fd1379663

[38] Frederiksen S. District Heating and Cooling. 1st ed. Lund: Studentlitteratur; 2013. 586 p

[39] De Souza Silva JL, Moreira HS, Dos Reis MVG, Barros TADS, Villalva MG. Theoretical and behavioral analysis of power optimizers for grid-connected photovoltaic systems. Energy Reports. 2022;**8**:10154-10167

[40] de Souza Silva JL, Mahmoudi E, Carvalho RRM, dos Santos Barros TA. Classification of anomalies in photovoltaic systems using supervised machine learning techniques and real data. Energy Reports. 2024;**11**:4642-4656

[41] Li X, Wang Q, Xu C, Wu Y, Li L. Survey of lithium-ion battery anomaly detection methods in electric vehicles. IEEE Transactions on Transportation Electrification. 2025;**11**(1):4189-4201

[42] Neumayer M, Stecher D, Grimm S, Maier A, Bücker D, Schmidt J. Fault and anomaly detection in district heating substations: A survey on methodology and data sets. Energy. 2023;**276**:127569. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85153487782&doi=10.1016%2fj.energy.2023.127569&partnerID=40&md5=1d2ea8bd8b1b3e2b6f1368acf4f37c74

[43] Bai M, Liu J, Chai J, Zhao X, Yu D. Anomaly detection of gas turbines based on normal pattern extraction. Applied Thermal Engineering. 2020;**166**:114664. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076598604&doi=10.1016%2fj.applthermaleng.2019.114664&partnerID=40&md5=03fa9a404a482d8fba9d6120f1e9db92

[44] Chen H, Liu H, Chu X, Liu Q, Xue D. Anomaly detection and critical SCADA parameters identification for wind turbines based on LSTM-AE neural network. Renewable Energy. 2021;**172**:829-840

[45] Jiang J, Zhang R, Wu Y, Chang C, Jiang Y. A fault diagnosis method for electric vehicle power lithium battery based on wavelet packet decomposition. Journal of Energy Storage. 2022;**56**:105909. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140489562&doi=10.1016%2fj.est.2022.105909&partnerID=40&md5=c239fcbda50e0c3de3a3b06baa3ca2b7

[46] Taghezouit B, Harrou F, Sun Y, Arab AH, Larbes C. Multivariate statistical monitoring of photovoltaic plant operation. Energy Conversion and Management. 2020;**205**:112317. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077460624&doi=10.1016%2fj.enconman.2019.112317&partnerID=40&md5=9c6c472583117a5fcebf095b1742890a

[47] Wang Y, Han X, Xu X, Pan Y, Dai F, Zou D, et al. A comprehensive data-driven assessment scheme for power battery of large-scale electric vehicles in cloud platform. Journal of Energy Storage. 2023;**64**:107210. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85151028561&doi=10.1016%2fj.est.2023.107210&partnerID=40&md5=54d4705677c99be21a4a3a62dc3e5483

[48] Algikar P, Xu Y, Yarahmadi S, Mili L. A robust data-driven process modeling applied to time-series stochastic power flow. IEEE Transactions on Power Systems. 2024;**39**(1):693-705

[49] Abolmasoumi AH, Farahani A, Mili L. Robust particle filter design with an application to power system state estimation. IEEE Transactions on Power Systems. 2024;**39**(1):1810-1821

[50] Forootani A, Rastegar M, Sami A. Short-term individual residential load forecasting using an enhanced machine learning-based approach based on a feature engineering framework: A comparative study with deep learning methods. Electric Power Systems Research. 2022;**210**:108119. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85133966897&doi=10.1016%2fj.epsr.2022.108119&partnerID=40&md5=30704793ef9e235244c2995e2d615e31

[51] Zhang S, Robinson E, Basu M. Wind turbine condition monitoring based on three fitted performance curves. Wind Energy. 2024;**27**(5):429-446

[52] Obermayr M, Riess C, Wilde J. A novel online 4-point rainflow counting algorithm for power electronics. Microelectronics Reliability. 2021;**120**:114112

[53] Yin S, Yang H, Xu K, Zhu C, Zhang S, Liu G. Dynamic real–time abnormal energy consumption detection and energy efficiency optimization analysis considering uncertainty. Applied Energy. 2022;**307**:118314. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85120745388&doi=10.1016%2fj.apenergy.2021.118314&partnerID=40&md5=bec14803ea2627ec4d8c06762c664c8e

[54] Jin F, Wu H, Liu Y, Zhao J, Wang W. Varying-scale HCA-DBSCAN-based anomaly detection method for multi-dimensional energy data in steel industry. Information Sciences. 2023;**647**(C):119479. DOI: 10.1016/j.ins.2023.119479

[55] Yao Q, Hu Y, Liu J, Zhao T, Qi X, Sun S. Power curve modeling for wind turbine using hybrid-driven outlier detection method. Journal of Modern Power Systems and Clean Energy. 2023;**11**(4):1115-1125

[56] Elnour M, Meskin N, Khan K, Jain R. Application of data-driven attack detection framework for secure operation in smart buildings. Sustainable Cities and Society. 2021;**69**:102816. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102971255&doi=10.1016%2fj.scs.2021.102816&partnerID=40&md5=48c224deae535fddfedb4cd21a3f7565

[57] Li T, Liu X, Lin Z, Morrison R. Ensemble offshore wind turbine power curve modelling—An integration of isolation forest, fast radial basis function neural network, and metaheuristic algorithm. Energy. 2022;**239**:122340. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117364439&doi=10.1016%2fj.energy.2021.122340&partnerID=40&md5=d37af9463db3a3e12a11ce7fc6e4592f

[58] Kong J, Jiang W, Tian Q, Jiang M, Liu T. Anomaly detection based on joint spatio-temporal learning for building electricity consumption.

Applied Energy. 2023;**334**:120635. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146457701&doi=10.1016%2fj.apenergy.2022.120635&partnerID=40&md5=e50a22ff89197a29aa5df898a64734c5

[59] Gao H, Yang D, Cai G, Chen Z, Ma J, Wang L, et al. Machine learning-based reliability improvement of ambient mode extraction for smart grid utilizing isolation forest. IEEE Transactions on Power Systems. 2023;**38**(5):4752-4760

[60] Asefi S, Mitrovic M, Ćetenović D, Levi V, Gryazina E, Terzija V. Anomaly detection and classification in power system state estimation: Combining model-based and data-driven methods. Sustainable Energy, Grids and Networks. 2023;**35**:101116. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85166961930&doi=10.1016%2fj.segan.2023.101116&partnerID=40&md5=856d8ba5448d11595cf0fa323e84a0b0

[61] Diao W, Naqvi IH, Pecht M. Early detection of anomalous degradation behavior in lithium-ion batteries. Journal of Energy Storage. 2020;**32**:101710. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089036032&doi=10.1016%2fj.est.2020.101710&partnerID=40&md5=c51a8199f4c6e012e1406491b203bc16

[62] Dao PB. On Wilcoxon rank sum test for condition monitoring and fault detection of wind turbines. Applied Energy. 2022;**318**:119209

[63] Dao PB, Barszcz T, Staszewski WJ. Anomaly detection of wind turbines based on stationarity analysis of SCADA data. Renewable Energy. 2024;**232**:121076. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85199995353&doi=10.1016%2fj.renene.2024.121076&partnerID=40&md5=e450c7bffbe0cc0ac8ebd77964cc9ae6

[64] Livera A, Theristis M, Micheli L, Stein JS, Georghiou GE. Failure diagnosis and trend-based performance losses routines for the detection and classification of incidents in large-scale photovoltaic systems. Progress in Photovoltaics: Research and Applications. 2022;**30**(8):921-937

[65] Ohunakin OS, Henry EU, Matthew OJ, Ezekiel VU, Adelekan DS, Oyeniran AT. Conditional monitoring and fault detection of wind turbines based on Kolmogorov–Smirnov non-parametric test. Energy Reports. 2024;**11**:2577-2591

[66] Qiu Y, Peng P, Jiang F. Improvement of local outlier factor algorithms for lithium-ion battery fault diagnosis. Journal of Energy Storage. 2024;**98**:113100. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85199787191&doi=10.1016%2fj.est.2024.113100&partnerID=40&md5=a10ba01e78c6d65ede66a6cf26f564bb

[67] Yuan H, Cui N, Li C, Cui Z, Chang L. Early stage internal short circuit fault diagnosis for lithium-ion batteries based on local-outlier detection. Journal of Energy Storage. 2023;**57**:106196. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85143536926&doi=10.1016%2fj.est.2022.106196&partnerID=40&md5=8e6b2dd41356291bb1fa93c0163a1276

[68] Ristic B. Particle Filters for Random Set Models. New York, NY: Springer New York; 2013. Available from: https://link.springer.com/10.1007/978-1-4614-6316-0

[69] Wang Y, Hu Q, Li L, Foley AM, Srinivasan D. Approaches to wind power

curve modeling: A review and discussion. Renewable and Sustainable Energy Reviews. 2019;**116**:109422

[70] Khezami N, Benhadj Braiek N, Guillaud X. Wind turbine power tracking using an improved multimodel quadratic approach. ISA Transactions. 2010;**49**(3):326-334

[71] Turnbull A, Carroll J, McDonald A. Combining SCADA and vibration data into a single anomaly detection model to predict wind turbine component failure. Wind Energy. 2021;**24**(3):197-211

[72] Morrison R, Liu X, Lin Z. Anomaly detection in wind turbine SCADA data for power curve cleaning. Renewable Energy. 2022;**184**:473-486

[73] Khan PW, Byun YC. Detecting anomaly classification using PCA-Kmeans and ensembled classifier for wind turbines. IEEE Open Access Journal of Power and Energy. 2024;**11**:349-361

[74] Trizoglou P, Liu X, Lin Z. Fault detection by an ensemble framework of Extreme Gradient Boosting (XGBoost) in the operation of offshore wind turbines. Renewable Energy. 2021;**179**:945-962

[75] Mascali L, Schiera DS, Eiraudo S, Barbierato L, Giannantonio R, Patti E, et al. A machine learning-based anomaly detection framework for building electricity consumption data. Sustainable Energy, Grids and Networks. 2023;**36**. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85175365401&doi=10.1016%2fj.segan.2023.101194&partnerID=40&md5=9394e4198b60c8ca6e5a5f6429731f7f

[76] Lei L, Wu B, Fang X, Chen L, Wu H, Liu W. A dynamic anomaly detection method of building energy consumption based on data mining technology. Energy. 2023;**263**:125575. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85139299674&doi=10.1016%2fj.energy.2022.125575&partnerID=40&md5=6481fd85df259a6aeabc1007bac94dd5

[77] Zhang L, Guo J, Lin P, Tiong RLK. Detecting energy consumption anomalies with dynamic adaptive encoder-decoder deep learning networks. Renewable and Sustainable Energy Reviews. 2025;**207**:114975. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85206268172&doi=10.1016%2fj.rser.2024.114975&partnerID=40&md5=c158cbd6ab47c830b2d22b064da4be98

[78] Qi R, Li Q, Luo Z, Zheng J, Shao S. Deep semi-supervised electricity theft detection in AMI for sustainable and secure smart grids. Sustainable Energy, Grids and Networks. 2023;**36**:101219. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85177026911&doi=10.1016%2fj.segan.2023.101219&partnerID=40&md5=8d65d3cb86d60e3ad933d6ba75e663ac

[79] Zhu Y, Zhu C, Tan J, Tan Y, Rao L. Anomaly detection and condition monitoring of wind turbine gearbox based on LSTM-FS and transfer learning. Renewable Energy. 2022;**189**:90-103

[80] Tobar A, Flores M, Castillo-Páez S, Naya S, Zaragoza S, Tarrío-Saavedra J. Bootstrap-LOCI data mining methodology for anomaly detection in buildings energy efficiency. Energy Reports. 2023;**10**:244-254

[81] Choubey M, Chaurasiya RK, Yadav JS. Contrastive learning for efficient anomaly detection in electricity load data. Sustainable Energy, Grids and Networks. 2025;**42**:101639. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85217280793&doi=10.1016%2fj.segan.20

25.101639&partnerID=40&md5=078584789d3b60aaa093606d8b3563b0

[82] Cai Q, Li P, Zhao Z, Wang R. Dynamic electricity theft behavior analysis based on active learning and incremental learning in new power systems. International Journal of Electrical Power & Energy Systems. 2024;**162**:110309. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85207691006&doi=10.1016%2fj.ijepes.2024.110309&partnerID=40&md5=2458a2ae9bc0e4de6a959c5bcb91407e

[83] Yang J, Chen X, Zhao C. Toward the ensemble consistency: Condition-driven ensemble balance representation learning and nonstationary anomaly detection for battery energy storage system. Applied Energy. 2025;**381**:125160. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85212562453&doi=10.1016%2fj.apenergy.2024.125160&partnerID=40&md5=ab8023cc5f01dbaaa539d36e7b85a151

[84] Shi Y, He W, Zhao J, Hu A, Pan J, Wang H, et al. Expected output calculation based on inverse distance weighting and its application in anomaly detection of distributed photovoltaic power stations. Journal of Cleaner Production. 2020;**253**:119965. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077397181&doi=10.1016%2fj.jclepro.2020.119965&partnerID=40&md5=b6c5565334b7034efdece9d629fc1f3f

[85] Hermans C, Al Koussa J, Van Oevelen T, Vanhoudt D. Fault detection for district heating substations: Beyond three-sigma approaches. Smart Energy. 2024;**16**:100159. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205483172&doi=10.1016%2fj.segy.2024.100159&partnerID=40&md5=f07ae8242fb01830ae209d18e4e7d738

[86] Lin F, Zhang X, Li X, Sun C, Zsurzsan G, Cai W, et al. AI-based design with data trimming for hybrid phase shift modulation for minimum-current-stress dual active bridge converter. IEEE Journal of Emerging and Selected Topics in Power Electronics. 2024;**12**(2):2268-2280

[87] Li S, Pandey A, Hooi B, Faloutsos C, Pileggi L. Dynamic graph-based anomaly detection in the electrical grid. IEEE Transactions on Power Systems. 2022;**37**(5):3408-3422

[88] Castangia M, Sappa R, Girmay AA, Camarda C, Macii E, Patti E. Anomaly detection on household appliances based on variational autoencoders. Sustainable Energy, Grids and Networks. 2022;**32**. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85145802031&doi=10.1016%2fj.segan.2022.100823&partnerID=40&md5=cc63c3df3de9be1a900a41149fb9ae23

[89] Movahed P, Taheri S, Razban A. A bi-level data-driven framework for fault-detection and diagnosis of HVAC systems. Applied Energy. 2023;**339**:120948. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85150300115&doi=10.1016%2fj.apenergy.2023.120948&partnerID=40&md5=bafeb38b4d246268d47f166bed4ae03d

[90] Jin F, Zhao J, Han Z, Wang W. A joint scheduling method for multiple byproduct gases in steel industry. Control Engineering Practice. 2018;**80**:174-184

[91] Wang T, Zhao J, Liu Q, Wang W. Granular-based multilayer spatiotemporal network with control gates for energy prediction of steel industry. IEEE Transactions on Instrumentation and Measurement. 2021;**70**:1-12

[92] Souza Santos IA, De Medeiros Santos VR, Dos Reis Lima W, Da

Silva AL, Maia BT, De Oliveira JR. Slag splashing: Simulation and analysis of the slags conditions. Journal of Materials Research and Technology. 2019;**8**(6):6173-6176

[93] Li Y, Shen X. Anomaly detection and classification method for wind speed data of wind turbines using spatiotemporal dependency structure. IEEE Transactions on Sustainable Energy. 2023;**14**(4):2417-2431

[94] Gao B, Kong X, Li S, Chen Y, Zhang X, Liu Z, et al. Enhancing anomaly detection accuracy and interpretability in low-quality and class imbalanced data: A comprehensive approach. Applied Energy. 2024;**353**:122157. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85175533259&doi=10.1016%2fj.apenergy.2023.122157&partnerID=40&md5=21851b4184f0669e48a62752e614cea5

[95] Huyghues-Beaufond N, Tindemans S, Falugi P, Sun M, Strbac G. Robust and automatic data cleansing method for short-term load forecasting of distribution feeders. Applied Energy. 2020;**261**:122157. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077647677&doi=10.1016%2fj.apenergy.2019.114405&partnerID=40&md5=4bbcdc278f19b9a0f16375aaffec7ce1

[96] Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: A survey on methods and metrics. Electronics. 2019;**8**(8):832

[97] Li Z, Zhu Y, Van Leeuwen M. A survey on explainable anomaly detection. ACM Transactions on Knowledge Discovery from Data. 2023;**18**(1):1-54. DOI: 10.1145/3609333

[98] Vinh NX, Chan J, Romano S, Bailey J, Leckie C, Ramamohanarao K, et al. Discovering outlying aspects in large datasets. Data Mining and Knowledge Discovery. 2016;**30**(6):1520-1555

[99] Amarasinghe K, Kenney K, Manic M. Toward explainable deep neural network based anomaly detection. In: 11th International Conference on Human System Interaction (HSI). Gdansk, Poland: IEEE; 2018. pp. 311-317. Available from: https://ieeexplore.ieee.org/document/8430788/

[100] Kauffmann J, Ruff L, Montavon G, Müller KR. The clever Hans effect in anomaly detection. arXiv. 2020. Available from: https://arxiv.org/abs/2006.10609

[101] Antwarg L, Miller RM, Shapira B, Rokach L. Explaining anomalies detected by autoencoders using SHAP. arXiv. 2019. Available from: https://arxiv.org/abs/1903.02407

[102] Yao R, Wang N, Ke W, Liu Z, Yan Z, Sheng X. Electricity theft detection in incremental scenario: A novel semi-supervised approach based on hybrid replay strategy. IEEE Transactions on Instrumentation and Measurement. 2023;**72**:1-12

[103] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Saarbrucken: IEEE; 2016. pp. 372-387. Available from: http://ieeexplore.ieee.org/document/7467366/

[104] Lemley J, Bazrafkan S, Corcoran P. Deep learning for consumer devices and services: Pushing the limits for machine learning, artificial intelligence, and computer vision. IEEE Consumer Electronics Magazine. 2017;**6**(2):48-56

Chapter 2

Fraud Detection in E-Commerce: A Systematic Review of Transaction Risk Prevention

Susie Xi Rao, Jiawei Jiang, Zhichao Han and Hang Yin

Abstract

We provide a systematic review of the endeavors of e-commerce companies in combating transaction risks that involve buyers, sellers, items, and transactions. There has been a paradigm shift from rule-based systems to simple machine learning-based systems to deep learning-based systems. This transformation has also involved multi-modal data engineering efforts like rule extraction, feature engineering on text and image, graph-structured abstraction, etc. In this review, we not only reflect on the shifts in data and systems but also the role of human experts, as well as the infra-structure support for such shifts, which are oftentimes neglected in previous review articles. The key conclusions from this review are three. First, there has been an increasing focus on multimodal data engineering efforts, explainability, and human-in-the-loop systems. Second, despite certain contributions to the online scalability of fraud detection systems, this topic has remained understudied. Third, newer research trends are on federated learning and adversarial machine learning, reinforcement learning, large language models, and their applicability and feasibility to integrate into the existing e-commerce fraud detection systems.

Keywords: fraud detection, e-commerce, transaction risk, artificial intelligence, review

1. Introduction

Anomaly detection plays a crucial role in modern e-commerce platforms to mitigate transaction risks involving buyers, sellers, items, and transactions. This chapter systematically reviews the evolution from rule-based systems to machine learning and deep learning-based methods, along with the infrastructure and human expertise needed for their deployment.

Traditional rule-based fraud detection systems rely on predefined heuristics and domain expertise to flag suspicious activities. While effective for well-known fraud patterns, these systems struggle with adapting to evolving fraudulent tactics. The advent of machine learning and deep learning approaches has significantly improved fraud detection by leveraging large-scale transactional data, behavioral patterns, and graph-based representations. Graph-based anomaly detection has emerged as a

powerful tool for capturing relational structures in fraud detection, which strengthens the ability to model complex interactions between entities such as users, devices, and transactions.

This chapter explores various aspects of fraud detection in e-commerce, including feature engineering, which plays a crucial role in constructing informative representations for anomaly detection models. Explainability and human-in-the-loop systems highlight the need for transparency and interpretability in AI-driven fraud detection, ensuring regulatory compliance and building trust among fraud analysts. Scalability challenges are discussed in the context of real-time fraud detection, where millions of transactions must be processed efficiently.

Additionally, we examine federated learning, which enables multiple e-commerce platforms to collaboratively train fraud detection models while preserving data privacy. Adversarial machine learning is another emerging area, addressing fraudsters' ability to manipulate detection systems by generating synthetic transaction patterns or evading anomaly detection models. We also critically assess the feasibility to utilize large language models in facilitating fraud detection workloads.

As fraud tactics become increasingly sophisticated, the integration of advanced graph-based learning, adversarial training, large language models, and privacy-preserving techniques will be essential in strengthening fraud detection capabilities across diverse e-commerce environments. This chapter provides a comprehensive review of these advancements, identifying key challenges and future research directions in anomaly detection for e-commerce fraud prevention.

2. Fundamentals of anomaly detection in e-commerce

Anomalies refer to instances that deviate significantly from normal patterns. These anomalies can manifest as fraudulent transactions, fake reviews, identity theft, or collusive seller-buyer activities. The primary challenge in anomaly detection lies in differentiating between genuine outliers and fraudulent activities, as well as handling evolving fraud patterns that adapt to detection mechanisms. This differentiation is usually contingent on automatically generated metrics or/and human judgment and varies according to use cases.

Early fraud detection systems relied on predefined rules and expert knowledge. These systems flagged transactions based on thresholds for parameters such as transaction amount, frequency, and location [1]. While effective for detecting known fraud patterns, rule-based systems struggle with adaptive fraud strategies. Additionally, they require constant updates from fraud analysts to stay relevant against emerging threats.

However, rule extraction techniques are not obsolete. They serve as basic building blocks for advanced fraud detection systems, some of which rely on predefined rules to construct their graphs, especially on criteria of node connections [2, 3]; others utilize the rules in post-detection to explain the system output [4–6]. The newly extracted rules will be augmented to the existing rule-based engines that have already been implemented on the platform.

Machine learning techniques offer more flexibility, leveraging feature engineering and (mostly) supervised learning models to detect fraudulent transactions [7]. Simple algorithms such as decision trees, random forests, and support vector machines are adopted to analyze historical transaction data to identify suspicious patterns. In practice, tree-based models such as gradient boosting decision tree (GBDT) are extremely

powerful in large-scale e-commerce applications [5, 8]. These models reduce reliance on manual rule updates by learning from labeled datasets. However, their effectiveness depends on high-quality labeled data and well-engineered features.

Advanced fraud detection systems integrate deep learning models and graph-based representations, including graph neural networks (GNNs), to enhance accuracy and scalability. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been used to detect fraud based on sequential transaction data [5, 6, 9–12]. Graph-based models analyze relationships between users, transactions, and payment methods in scenarios such as anti-money laundering [13–18], credit card fraud detection [19], user onboarding [20], transaction risks [2–4, 8, 21–25], and risky accounts [25–28].

Graph structures help detect fraud rings and suspicious relationships among users, transactions, and payment details. By constructing transaction/user/device networks, anomalies can be detected through graph analytics techniques such as community detection and centrality measures. For example, fraudsters often use multiple fake accounts to artificially boost product ratings or conduct money laundering activities, which can be identified through dense subgraph patterns.

3. Case studies in e-commerce fraud detection

Now we zoom in on the aspects that are of special focus in both academia and industry on state-of-the-art fraud detection performance. We provide (**Table 1**) as a reference for readers to retrieve the relevant literature items.

We survey the publications on e-commerce fraud detection from the top-ranked venues on machine learning, deep learning, data mining, and graph representation learning between the years 2017 and 2024, including but not limited to AAAI, CIKM, Expert Systems with Applications, ICDE, ICDM, ICLR, IEEE venues, IJCAI, KDD, NeurIPS, TKDE, VLDB, WSDM, and WWW.[1] There has been a surge in increasing usage of advanced machine learning techniques and deep learning techniques with the further development of sufficient infrastructures and the abundance of data availability. Since the year 2018, there has been an increase in using graph representation learning (with GNNs as examples) to further advance pattern detection based on the existing infrastructures.

Table 1 summarizes key studies on fraud detection in this review. The *Graph* column specifies the type of graph used, such as transaction networks or account networks. The *Construction Method* describes how each graph was built, following the taxonomy in Ref. [36], including sources such as external knowledge bases, raw data, similarity metrics, or multimodal content. *Graph Type* characterizes the graph's structural properties, such as bipartite, homogeneous, heterogeneous, dynamic, or multimodal. The *Data* column lists the dataset(s) used and the public availability.[2] In the column *Application*, we specify the fraud detection domain, such as anti-money laundering or risky account detection. The *Method* column outlines the modeling

[1] For a more comprehensive survey of the fraud detection landscape that includes over 180 publications, we refer the readers to Rao et al. [35]. Another source that is actively curated is https://github.com/safe-graph/graph-fraud-detection-papers.

[2] Since many publications do not document dataset links, or provided links are outdated, the authors have made an effort to locate accessible sources.

Graph	Construction method	Graph type	Data	Application	Method	Reference	Code availability	Year
Transaction graph	Data (usage)	Homogeneous	Elliptic	Anti-money laundering	GNN	[18]	No	2019
Bank account graph	Data (money transfer)	Heterogeneous Multi-partite	CBank (private) CFD	Anti-money laundering	Non-GNN (FlowScope)	[29]	FlowScope	2020
User-transaction graph	Data (transactions)	Homogeneous	WeBank-small/-medium/-large (private) OTC Alpha Elliptic	Anti-money laundering	GNN (Diga)	[14]	No	2023
Transaction graph	Data (related transactions)	Homogeneous	Alipay data (private)	Anti-money laundering	GNN	[15]	No	2023
User-transaction graph	Data (transactions)	Dynamic Homogeneous	UnionPay data (private)	Anti-money laundering	GNN	[30]	No	2023
Transaction graph	Data (related transactions)	Homogeneous	Bitcoin transaction dataset	Anti-money laundering	GNN	[16]	No	2023
User-transaction graph	Data (transactions)	Homogeneous	AMLSim Elliptic	Anti-money laundering	GNN	[17]	No	2023
Transaction graphs	Data (relations)	Dynamic Heterogeneous	HI LI (small, medium, large)	Anti-money laundering	GNN (Multi-GNN)	[13]	Multi-GNN	2024
User login behavior graph	Data (relations)	Heterogeneous	Jingdong e-commerce (private)	Collective fraud	Non-GNN (statistic or other anomaly detection algorithms, HGsuspector)	[31]	No	2018
Transaction graph	Data (usage)	Heterogeneous	Vesta Sparkov	Credit card fraud detection	GNN (FIW-GNN)	[19]	No	2023
Transaction graph	Data (relations)	Heterogeneous	ELECTRONIC ARTS data (private)	Fraudulent transactions	Non-GNN (predefined meta-paths, HitFraud)	[32]	No	2017
Transaction-user graph	Data (usage)	Dynamic Heterogeneous Multimodal	Alipay data (private)	Fraudulent transactions	RNN	[11]	No	2019

Graph	Construction method	Graph type	Data	Application	Method	Reference	Code availability	Year
User-attribute graph	Data (relations)	Heterogeneous	Ant Credit Pay dataset (private)	Fraudulent transactions Fraudulent users	Non-GNN (predefined meta-paths, HACUD)	[33]	DGFraud	2019
Transaction graph	Data (usage)	Dynamic Homogeneous	Alipay data (private)	Fraudulent transactions	Non-GNN (sequence-based)	[34]	No	2019
User behavior sequence	Data (usage)	Dynamic	E-commerce company data from four countries (private)	Fraudulent transactions	Non-GNN (sequence-based, hierarchical, transfer learning)	[9]	No	2020
Transaction-attribute graph	Data (usage)	Heterogeneous	*eBay-small* (per request) *eBay-large/-xlarge* (private)	Fraudulent transactions	GNN (transformer)	[24]	xFraud	2021
Transaction graph	Data (rules)	Homogeneous	BTFSD (private)	Fraudulent transactions	GNN (knowledge-guided, graph, Know-GNN)	[3]	No	2021
Transaction graph	Data (rules)	Homogeneous	BTFSD (private) IEEE-CIS Fraud Detection YelpCHi	Fraudulent transactions	GNN (knowledge-guided, rules, KS-GNN)	[2]	No	2021
Intention-transaction graph	Data (usage)	Heterogeneous	Alibaba graph (private)	Fraudulent transactions	GNN (IHGAT)	[4]	No	2021
Transaction graph	Data (usage)	Heterogeneous Multimodal	eBay (private)	Fraudulent transactions	Bi-LSTM GPU-powered HDBSCAN clustering (FinDeepBehaviorCluster)	[6]	No	2021
Transaction-attribute graph	Data (usage)	Heterogeneous Dynamic	eBay (private)	Fraudulent transactions	GNN (BRIGHT)	[8]	No	2022
Transaction graph	Data (usage)	Heterogeneous Multimodal	eBay (private)	Fraudulent transactions	Bi-LSTM Transformer (BehaviorClustering)	[5]	No	2022
Transaction graph	Data (relations)	Dynamic Heterogeneous	PR01 (private) Orange Finance Company (upon request)	Fraudulent transactions	GNN (Transformer, STA-GT)	[23]	No	2023
Transaction graph	Data (usage)	Heterogeneous Multimodal	eBay (private)	Fraudulent transactions	Transformer (MMBT)	[12]	No	2024

Graph	Construction method	Graph type	Data	Application	Method	Reference	Code availability	Year
User-behavior graph Review graph Transaction graph	Data (relations, usage)	Heterogeneous Homogeneous	Reddit Weibo Amazon YelpChi Tolokers Questions T-Finance Elliptic DGraph-Fin T-Social	Graph anomaly detection	GNN (GADBench)	[21]	GADBench	2023
User review graph User-transaction graph	Data (homophily and heterophily edges)	Heterogeneous	YelpCHi Amazon T-Finance	Graph anomaly detection	GNN (spectral, SEC-GFD)	[22]	No	2024
Account graph	Data (usage)	Bipartite Dynamic	Amazon account activity data (private)	Risky account detection	GNN	[27]	No	2018
Device-account graph	Data (usage)	Heterogeneous	Alipay data (private)	Risky account detection	GNN (GEM)	[26]	GEM	2018
User-account graph	Data (usage)	Homogeneous	Pubmed BlogCatalog (no feature) BlogCatalog (128-dimension feature) PPI Alipay (private)	Risky account detection	GNN (breadth and depth, GeniePath)	[28]	GeniePath	2019
User-behavior graph	Data (relations)	Dynamic Homogeneous	Elliptic T-Finance T-Social YelpCHi Amazon	Suspicious account/ transaction	GNN (DGA-GNN)	[25]	DGA-GNN	2024

Table 1.
Summary of related works in fraud detection, sorted alphabetically first by the column Application and then by the column Year.

techniques in each study, and *Reference* provides citations for further reading. Finally, we survey the code availability and publication year in the last two columns.

3.1 Feature engineering

Feature engineering is a crucial step in improving fraud detection performance. E-commerce platforms extract meaningful features from textual descriptions, visual product images, and behavioral data such as browsing history and purchase frequency. For instance, fraudulent users can be detected by analyzing abnormal mouse trajectories [12], while suspicious accounts may be flagged based on IP address discrepancies and login patterns [27].

The choice of features influences the ability of the downstream models to distinguish fraudulent patterns from legitimate behaviors. Existing works employ a variety of feature engineering techniques tailored to the underlying data characteristics, graph structure of attributes, and anomaly detection tasks.

3.1.1 Feature extraction and graph construction

Fraud detection studies leverage different types of graphs, such as transaction graphs, user-transaction graphs, and account graphs. These graphs are constructed based on various criteria, including transaction relations [13], account behaviors [27], and financial interactions [16]. Feature extraction methods commonly include node embeddings, edge attributes, and temporal patterns, with several studies using predefined meta-paths [32, 33] to enhance feature representations.

3.1.2 Structural and relational features

To capture relational and structural dependencies, previous studies derive features based on node connectivity, degree distributions, and centrality measures. For instance, in anti-money laundering applications, Erik Altman et al. [13] incorporate multi-type relationships to better represent complex financial ecosystems. In credit card fraud detection, Yan et al. [19] extract statistical patterns from transaction sequences to identify suspicious activities.

3.1.3 Temporal and behavioral features

Dynamic features are increasingly adopted to model evolving fraudulent behaviors. Studies such as Cao et al., Tian et al., and Cheng et al. [23, 30, 34] analyze transaction sequences over time, leveraging time-based embeddings to improve detection accuracy. Sequence-based methods, including behavioral clustering [5] and hierarchical transfer learning [9], further enhance fraud detection by capturing sequential user activity patterns.

3.1.4 Multimodal and knowledge-guided features

To improve feature richness, several approaches integrate multimodal data sources. For instance, transaction-attribute graphs combine user behaviors with additional contextual information, such as transaction intentions [4]. Knowledge-guided techniques, such as rule-based graph construction [2, 3], embed domain knowledge into graph representations, improving interpretability and model robustness.

3.1.5 Graph neural networks and alternative approaches

While many recent approaches utilize GNN-based feature extraction [17, 21], non-GNN methods, such as statistical anomaly detection [31] and flow-based models [29], remain relevant in certain fraud detection scenarios. The choice of feature engineering techniques is often driven by the nature of the data, the computational complexity of graph models, and the need for explainability.

Overall, feature engineering in anomaly detection continues to evolve, with a growing emphasis on dynamic, multimodal, and knowledge-guided representations. Future work may focus on improving feature selection strategies and developing more interpretable representations for fraud detection applications, for instance, with the help of large language models.

3.2 Integration of multimodal data for fraud detection

E-commerce platforms integrate multiple data sources, such as transaction logs, clickstream data, image-based product verification, and external knowledge bases, to improve fraud detection capabilities. Behavioral sequence embeddings from user activity logs can reveal unusual navigation patterns [12], while product image analysis [37] and grounded knowledge bases [38] help detect counterfeit items and invalid connections. When combining these data sources, we have a holistic review of multiple sources that fraudsters can default. One example of combining multiple sources of input is the integration of sequence embeddings in fraud detection tasks.

Recent studies have explored various sequence embedding approaches for fraud detection, leveraging different techniques to enhance model performance and interpretability. For instance, Li et al. [11] proposed a time attention-based embedding method that incorporates dwell time to better capture user behavior patterns. Additionally, other works have focused on improving interpretability through intent-aware embeddings [4], multi-task learning for general behavior representation [6], and cross-domain feature embeddings [5]. These approaches collectively aim to enhance the accuracy and interpretability of fraud detection systems by capturing complex user behavior patterns and leveraging diverse data sources.

3.3 E-commerce companies as dataset contributors

In most of the contributions, e-commerce companies generate large-scale anonymized datasets, such as Amazon [2, 19, 25, 27], Alibaba [4, 11, 15, 21, 22, 26, 28, 33, 34], eBay [5, 6, 8, 12, 24], and Jingdong [31]. Due to privacy concerns, many datasets remain private or are available only upon request, limiting accessibility for broader research. As dataset availability significantly impacts research progress, future efforts may focus on developing more openly accessible (synthetic) benchmarks while ensuring data privacy and compliance with regulations.

4. Explainability and scalability in e-commerce fraud detection

4.1 Explainability and human-in-the-loop systems

Balancing automated detection with human expertise is essential. Explainable AI techniques [39], such as SHAP (Shapley Additive Explanations) and LIME (Local

Interpretable Model-Agnostic Explanations), help fraud analysts understand model decisions. Human-in-the-loop systems incorporate fraud investigators to review flagged transactions, reducing false positives and improving trust in AI-driven systems.

Understanding the rationale behind fraud detection predictions is crucial, particularly when interpreting why certain transactions are flagged as suspicious and how an algorithm contributes to this process. Typically, explainability methods can be divided into these categories [40]: rule-based [4–6], input-level [41–43], self-explanatory [44], and model-level explanations [45].

The importance of explainability extends to dynamic graphs, where fraud patterns evolve over time. Several works have focused on explaining temporal GNN models, including DGExplainer [46] for discrete-time graphs and TGNNExplainer [47] for continuous-time graphs. Given that financial fraud detection often relies on dynamic and heterogeneous graph structures, these techniques are crucial for identifying evolving fraudulent behaviors and providing justifications for decisions.

As the complexity of fraud detection models increases, improving interpretability remains a priority to ensure compliance with regulatory standards and foster trust in automated systems. Future research may focus on integrating human-in-the-loop feedback mechanisms with self-explainable models to enhance decision-making in real-world fraud detection applications.

4.2 Scalability and deployment challenges

Real-time fraud detection systems encounter notable obstacles in terms of scalability and deployment, especially when required to process millions of transactions every second. Ensuring scalability is vital for handling substantial financial datasets, including transaction graphs and patterns of user behavior, while maintaining rapid response times to detect fraudulent actions. Solutions like Apache Spark and PyTorch, along with cloud-based platforms, offer flexible computing resources to handle heavy traffic and facilitate effective anomaly detection. Nonetheless, the intensive computational demands of processing diverse and multimodal graphs continue to pose significant challenges.

Several studies have addressed scalability challenges by leveraging dynamic graph structures and efficient model architectures. For instance, transaction graphs from large-scale financial datasets [8, 34] require high-throughput processing to detect fraudulent behaviors in real-time. Dynamic graph models, such as those proposed by Huang et al. [48, 49], improve efficiency by capturing evolving patterns in streaming data. Additionally, optimizations like knowledge-guided rule embeddings [2] and behavior clustering techniques [5] help reduce processing overhead while maintaining detection accuracy.

Deploying fraud detection models introduces additional operational challenges, including latency constraints, regulatory compliance, and the need for continuous model retraining. E-commerce platforms must comply with data privacy regulations like GDPR and CCPA while ensuring secure transaction processing. Real-time fraud detection systems, such as those deployed by Alipay [30, 34] and eBay [8, 24], utilize dynamic and heterogeneous graph structures that require efficient processing pipelines.

Model retraining and adaptation are critical for maintaining the effectiveness of fraud detection systems. Fraud tactics continuously evolve, requiring models to be periodically updated with new labeled data. Dynamic graph-based approaches [8, 46]

and hybrid methods, such as rule-based embeddings [2] and behavior clustering techniques [5], enhance robustness by incorporating domain knowledge and real-time behavioral patterns. Federated learning and privacy-preserving techniques may offer solutions for training fraud detection models across multiple institutions while maintaining data confidentiality.

Future research should focus on hybrid solutions that combine distributed computing, graph compression, and lightweight neural architectures to improve scalability and real-time performance. Additionally, deployment strategies must prioritize computational efficiency, reduce bias in automated decisions, and integrate adaptive learning techniques to stay ahead of emerging fraudulent activities.

5. Future directions of fraud detection in e-commerce

5.1 Federated learning for fraud detection

Federated learning allows multiple e-commerce (sub-)platforms to collaboratively train fraud detection models without sharing raw data [50–52]. This decentralized learning approach improves data privacy while leveraging knowledge from different sources. By deploying federated learning, companies can strengthen fraud detection capabilities across multiple platforms without violating privacy regulations.

Recent studies have explored federated learning for fraud detection, particularly in handling diverse financial and transactional data. For instance, domain adaptation techniques [53] enhance the ability of federated models to generalize across different e-commerce environments, reducing performance degradation caused by distributional shifts. Cross-domain applications [54] further demonstrate the potential of federated learning in improving fraud detection across heterogeneous datasets, where transactional patterns may vary significantly between different platforms.

Handling multi-relational and heterogeneous data is another challenge in federated learning, particularly in fraud detection scenarios involving transaction graphs, user behavior networks, and account-device relationships. Federated approaches must efficiently aggregate knowledge while maintaining structural and relational integrity across multiple sources [53]. Adversarial learning and contrastive self-supervised techniques have been integrated into federated frameworks to enhance robustness against fraudulent adaptation [55].

However, federated learning in fraud detection also presents challenges such as communication overhead, non-IID (non-independent and identically distributed) data distributions, and adversarial attacks on decentralized models. Privacy-preserving techniques are important in mitigating risks associated with data leakage and model inversion attacks, such as secure multi-party computation and differential privacy. Future research may focus on optimizing federated aggregation strategies, developing adaptive fraud detection mechanisms for non-IID data, and improving the interpretability of federated fraud detection models to ensure regulatory compliance and trustworthiness.

5.2 Adversarial machine learning and fraudulent adaptation

Fraudsters continuously adapt to detection methods by leveraging adversarial techniques, such as generating synthetic transaction patterns to evade detection.

Techniques like adversarial training and anomaly-aware model regularization help defend against evolving fraudulent strategies.

Several studies have explored adversarial methods to enhance fraud detection and anomaly detection in dynamic and heterogeneous graph structures. For instance, adversarial graph-based models have been employed for anomaly detection in dynamic graphs [56], where evolving fraud tactics require adaptive learning techniques. In the context of fake news detection, adversarially trained GNNs, such as AA-HGNN [57] and DAGA-NN [58], improve robustness against misinformation-spreading strategies by capturing deceptive relationships in textual similarity and publisher networks.

Adversarial contrastive learning has also been explored for detecting fraudulent behaviors in claim-evidence graphs, as demonstrated in GETRAL [55], which enhances resilience against manipulation in misinformation detection. Additionally, adversarial defenses have been applied to both financial and e-commerce fraud detection. For example, ACD [59] integrates adversarial learning for anomaly detection in transaction graphs as anti-money laundering efforts. Similarly, adversarial techniques have been used to detect suspicious user behaviors in user-product and review networks, as seen in MO-GAA and PBAAD [60], which mitigate evasion attacks by fraudsters who manipulate online platforms.

Despite these advancements, adversarial attacks remain a significant challenge in fraud detection, particularly in continuously evolving environments. Future research may focus on developing more robust defenses, such as adaptive adversarial training, self-supervised anomaly-aware learning, and hybrid approaches that integrate knowledge graphs with adversarial robustness techniques. These methods can help improve the ability of fraud detection systems to counteract increasingly sophisticated fraudulent strategies while maintaining interpretability and scalability.

5.3 Reinforcement learning in fraud detection

Reinforcement learning (RL) has shown growing potential in addressing fraud detection challenges in dynamic fraud detection environments, particularly through RL integration with graph-based methods. Most of the contributions of RL-based methods have been developed for fraud detection domains such as fake reviews, fake news, and spam detection.

RL is used in GNN structures to select the optimal nodes, edges, and subgraphs. Dou et al. [61] introduce CARE-GNN, where RL is utilized to find the optimal amounts of neighbors to select in review graphs (e.g., Amazon, Yelp) for detecting fraud rings, while Yuan Gao et al. [62] employ RL to alleviate structural distribution shifts by aggregating homophilous neighbors. Recent advancements include REGAD [63], which uses RL to iteratively prune suspicious edges in product-purchase graphs, and RHGNN [64], which designs a reinforced neighbor selector for heterogeneous review-reviewer-product graphs while filtering out the camouflaged relationships.

RL has also been applied to competitive scenarios. Nash-Detect [65] models fraud detection as a minimax game between review spammers and spam detectors, and HP-KGAT [66] optimizes path-aware graph attention for fake news detection with subgraph reasoning. RL is appealing in unsupervised settings as the label scarcity is not the bottleneck of modeling. Typically, anomaly scores of nodes are computed, where a larger score means a higher abnormality. RAND [67] introduces RL-based neighborhood selection to amplify reliable signals in social and citation graphs, whereas RARE-GNN [68] learns anomaly-resistant GNNs without prior knowledge of ground truth in citation networks

like Cora, CiteSeer, and PubMed. Additionally, DiG-In-GNN [69] combines RL with contrastive learning on transaction graphs (e.g., T-Finance), and RoSGAS [70] integrates RL with self-supervised learning in a setting of multimodal social bot detection. Last but not least, RL is also used to address model interpretability in the literature. EXplainable Graph Neural Network (XGNN) [45] uses RL to explain Graph Convolutional Network (GCN) graph generation. These works collectively highlight the adaptability of RL in handling complex relational data, adversarial dynamics, and structural noise. However, challenges like computational overhead, reward design, and scalability remain.

Scalability challenges in real-time fraud detection—such as high latency in policy updates or computational bottlenecks during graph traversal—remain understudied. One possible solution is to utilize hierarchical reinforcement learning (HRL) frameworks where large graphs are divided into smaller subgraphs [71]. Another approach is to use lightweight deep-Q networks (DQN) with prioritized experience replay to speed up convergence [72]. Additionally, implementing edge computing solutions enables decentralized reinforcement learning inference [73]. Finally, distributed reinforcement learning approaches, such as federated graph RL [74], can improve scalability by distributing training across segmented graphs in parallel.

5.4 Large language models in fraud detection

The advent of large language models (LLMs), such as BERT [75], GPT-3 [76], and their successors have opened new opportunities for fraud detection in e-commerce. These models, pre-trained on vast corpora of text data, have demonstrated remarkable capabilities in understanding context, semantics, and even subtle patterns in unstructured data. Their application in fraud detection is particularly promising due to their ability to process and analyze multimodal data in a unified manner, including text, images, and transactional metadata.

One of the key strengths of LLMs lies in their ability to perform *contextual anomaly detection*. Unlike traditional models that rely on handcrafted features, LLMs can automatically infer complex relationships between entities (e.g., buyers, sellers, and items) and detect anomalies that deviate from normal patterns. For instance, LLMs can analyze product descriptions, reviews, and customer interactions to identify fraudulent listings or fake reviews with high precision [77–79].

Moreover, LLMs can be fine-tuned on domain-specific data to enhance their performance in fraud detection tasks. For example, fine-tuning BERT on e-commerce transaction data has been shown to improve the detection of fraudulent transactions by capturing subtle linguistic cues in transaction descriptions or customer communications [80]. Additionally, LLMs can be integrated with graph-based approaches to model relationships between entities, enabling the detection of sophisticated fraud schemes such as collusive fraud or money laundering.

However, challenges remain in deploying LLMs for fraud detection. The expenses associated with fine-tuning and inference, along with the requirement for extensive, high-quality labeled datasets, can be daunting. Furthermore, the interpretability of LLMs remains a concern, as their "black-box" nature makes it difficult to explain fraud detection decisions to stakeholders [81]. Future research should focus on developing more efficient and interpretable LLM architectures tailored for fraud detection, as well as exploring hybrid approaches that combine LLMs with traditional rule-based systems for enhanced robustness.

In conclusion, LLMs allow a transformative shift in fraud detection, which offers new capabilities for analyzing complex and multimodal data. As these models keep

advancing, they are expected to become integral to the upcoming generation of fraud detection systems, as long as issues concerning scalability, interpretability, and domain adaptation are resolved.

6. Conclusions

This chapter offers an extensive review of techniques for detecting anomalies aimed at preventing e-commerce fraud. We trace the progression from traditional rule-based systems to contemporary approaches that utilize machine learning and deep learning, emphasizing the significance of feature engineering, graph models, and the integration of multimodal data. Furthermore, we address scalability challenges, highlighting the need for real-time detection systems to handle millions of transactions per second with minimal delay and high precision. The importance of explainability and involving humans in the loop is underscored to ensure transparency and meet regulatory standards, thereby aiding fraud analysts in understanding model decisions comprehensively.

Emerging trends in fraud detection research include federated learning, adversarial learning, reinforcement learning, and the usage of large language models. Federated learning facilitates privacy-preserving collaboration between e-commerce platforms, allowing fraud detection models to generalize across different data sources without sharing raw information. Reinforcement learning allows fraud detection techniques to focus on informative nodes, edges, and/or subgraphs with enhanced past experiences from ground truth labels in a supervised setting or with dynamically learned policies in an unsupervised setting. Adversarial machine learning has become increasingly relevant as fraudsters develop sophisticated techniques to evade detection, requiring robust anomaly detection models that can resist adversarial attacks. Large language models allow new opportunities to efficiently integrate the existing pipelines on multimodal sources.

Despite these advancements, several challenges remain, including the need for adaptive learning strategies to handle evolving fraud patterns, efficient model deployment to optimize resource utilization in cloud-based environments, and scalable graph-based learning approaches to process large-scale transactional data. Future research may focus on integrating self-supervised learning for fraud detection, improving adversarial robustness in dynamic graph structures, and developing hybrid models that combine rule-based expertise with deep learning techniques.

Author details

Susie Xi Rao[1,2*†], Jiawei Jiang[3†], Zhichao Han[4] and Hang Yin[4]

1 Department of Management, Technology and Economics, ETH Zurich, Zurich, Switzerland

2 ETH AI Center, Zurich, Switzerland

3 Department of Computer Science, Wuhan University, Wuhan, China

4 eBay China, Shanghai, China

*Address all correspondence to: srao@ethz.ch

† These authors contributed equally.

IntechOpen

References

[1] del Mar Roldán-García M, García-Nieto J, Aldana-Montes JF. Enhancing semantic consistency in anti-fraud rule-based expert systems. Expert Systems with Applications. 2017;**90**:332-343

[2] Rao Y, Ren X, Duan C, Mi X, Cheng J, Yu C, et al. Knowledge-guided fraud detection using semi-supervised graph neural network. In: Web Information Systems Engineering–WISE 2021: 22nd International Conference on Web Information Systems Engineering, WISE 2021; Melbourne, VIC, Australia; October 26–29, 2021, Proceedings, Part I 22. Berlin, Heidelberg: Springer-Verlag; 2021. pp. 385-393

[3] Rao Y, Mi X, Duan C, Ren X, Cheng J, Chen Y, et al. Know-GNN: An explainable knowledge-guided graph neural network for fraud detection. In: International Conference on Neural Information Processing. Cham, Switzerland: Springer International Publishing; 2021. pp. 159-167

[4] Liu C, Sun L, Ao X, Feng J, He Q, Yang H. Intention-aware heterogeneous graph attention networks for fraud transactions detection. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery; 2021. pp. 3280-3288

[5] Yin H, Zhang Z, Wang Z, Özyurt Y, Liang W, Dong W, et al. Behavioral graph fraud detection in e-commerce. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW). New York City, U.S.: IEEE; 2022. pp. 1-8

[6] Min W, Liang W, Yin H, Wang Z, Li M, Lal A. Explainable deep behavioral sequence clustering for transaction fraud detection. In: The AAAI-21 Workshop on Knowledge Discovery from Unstructured Data in Financial Services. Washington DC, USA: AAAI Publisher; 2021. Available from: https://aaai-kdf.github.io/kdf2021/accepted_papers

[7] Mutemi A, Bacao F. E-commerce fraud detection based on machine learning techniques: Systematic literature review. Big Data Mining and Analytics. 2024;**7**(2):419-444

[8] Lu M, Han Z, Rao SX, Zhang Z, Zhao Y, Shan Y, et al. BRIGHT-Graph neural networks in real-time fraud detection. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. New York, NY, USA: Association for Computing Machinery; 2022. pp. 3342-3351

[9] Zhu Y, Xi D, Song B, Zhuang F, Chen S, Xi G, et al. Modeling users' behavior sequences with hierarchical explainable network for cross-domain fraud detection. In: Proceedings of The Web Conference 2020. New York, NY, USA: Association for Computing Machinery; 2020. pp. 928-938

[10] Zhang R, Zheng F, Min W. Sequential behavioral data processing using deep learning and the markov transition field in online fraud detection. arXiv preprint arXiv:1808.05329. 2018

[11] Li L, Liu Z, Chen C, Zhang Y-L, Zhou J, Li X. A time attention based fraud transaction detection framework. arXiv preprint arXiv:1912.11760. 2019

[12] Zhang Z, Yin H, Rao SX, Yan X, Wang Z, Liang W, et al. Identifying e-commerce fraud through user behavior data: Observations and insights. Data Science and Engineering. 2024;**10**:24-39,

[13] Altman E, Blanuša J, Von Niederhäusern L, Egressy B, Anghel A,

Atasu K. Realistic synthetic financial transactions for anti-money laundering models. Advances in Neural Information Processing Systems. 2024;**36**: 29851-29874

[14] Li X, Li Y, Mo X, Xiao H, Shen Y, Chen L. Diga: Guided diffusion model for graph recovery in anti-money laundering. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2023. pp. 4404-4413

[15] Chai Z, Yang Y, Dan J, Tian S, Meng C, Wang W, et al. Towards learning to discover money laundering sub-network in massive transaction network. AAAI. 2023;**37**(12):14153-14160

[16] Hyun W, Lee J, Suh B. Anti-money laundering in cryptocurrency via multi-relational graph neural network. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer-Verlag; 2023. pp. 118-130

[17] Karim MR, Hermsen F, Chala SA, de Perthuis P, Mandal A. Catch me if you can: Semi-supervised graph learning for spotting money laundering. arXiv preprint arXiv:2302.11880. 2023

[18] Weber M, Domeniconi G, Chen J, Weidele DKI, Bellei C, Robinson T, et al. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. arXiv preprint arXiv:1908.02591. 2019

[19] Yan K, Gao J, Matsypura D. FIW-GNN: A heterogeneous graph-based learning model for credit card fraud detection. In: 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA). New York City, U.S.: IEEE; 2023. pp. 1-10

[20] Rao SX, Lanfranchi C, Zhang S, Han Z, Zhang Z, Min W, et al. Modelling graph dynamics in fraud detection with "attention". arXiv preprint arXiv: 2204.10614. 2022

[21] Tang J, Hua F, Gao Z, Zhao P, Li J. Gadbench: Revisiting and benchmarking supervised graph anomaly detection. arXiv preprint arXiv:2306.12251. 2023

[22] Fan X, Wang N, Wu H, Wen X, Zhao X, Wan H. Revisiting graph-based fraud detection in sight of heterophily and spectrum. Proceedings of the AAAI Conference on Artificial Intelligence. 2024;**38**:9214-9222

[23] Tian Y, Liu G, Wang J, Zhou M. Transaction fraud detection via an adaptive graph neural network. arXiv preprint arXiv:2307.05633. 2023

[24] Rao SX, Zhang S, Han Z, Zhang Z, Min W, Chen Z, et al. xFraud: Explainable fraud transaction detection. Proceedings of the VLDB Endowment. 2021;**15**:427-436

[25] Duan M, Zheng T, Gao Y, Wang G, Feng Z, Wang X. DGA-GNN: Dynamic grouping aggregation gnn for fraud detection. Proceedings of the AAAI Conference on Artificial Intelligence. 2024;**38**:11820-11828

[26] Liu Z, Chen C, Yang X, Zhou J, Li X, Song L. Heterogeneous graph neural networks for malicious account detection. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. New York, NY, USA: Association for Computing Machinery; 2018. pp. 2077-2085

[27] Ma J, Zhang D, Wang Y, Zhang Y, Pozdnoukhov A. GraphRAD: A graph-based risky account detection system. In: KDD 2018 Workshop. New York, NY, USA: Association for Computing Machinery; 2018

[28] Liu Z, Chen C, Li L, Zhou J, Li X, Song L, et al. GeniePath: Graph neural networks with adaptive receptive paths. Proceedings of the AAAI Conference on Artificial Intelligence. 2019;**33**: 4424-4431

[29] Li X, Liu S, Li Z, Han X, Shi C, Hooi B, et al. Flowscope: Spotting money laundering based on graphs. Proceedings of the AAAI Conference on Artificial Intelligence. 2020;**34**:4731-4738

[30] Cheng D, Ye Y, Xiang S, Ma Z, Zhang Y, Jiang C. Anti-money laundering by group-aware deep graph learning. IEEE Transactions on Knowledge and Data Engineering. 2023; **35**(12):12444-12457

[31] Li X, Zhang W, Xi J, Zhu H. HGsuspector: Scalable collective fraud detection in heterogeneous graphs. In: KDD 2018 Workshop. New York, NY, USA: Association for Computing Machinery; 2018

[32] Cao B, Mao M, Viidu S, Yu S, Philip. HitFraud: "A broad learning approach for collective fraud detection in heterogeneous information networks". In: 2017 IEEE International Conference on Data Mining (ICDM). New York City, U.S.: IEEE; 2017. pp. 769-774

[33] Hu B, Zhang Z, Shi C, Zhou J, Li X, Qi Y. Cash-out user detection based on attributed heterogeneous information network with a hierarchical attention mechanism. Proceedings of the AAAI Conference on Artificial Intelligence. 2019;**33**:946-953

[34] Cao S, Yang XX, Chen C, Zhou J, Li X, Qi Y. TitAnt: Online real-time transaction fraud detection in Ant Financial. Proceedings of the VLDB Endowment. 2019;**12**(12):2082-2093

[35] Rao S, X, Han Z, Yin H, Jiang J, Zhang Z, Zhao Y, et al. Fraud detection at eBay. Emerging Markets Review. 2025. Special Issue: Multi-Source Data-Driven Financial Fraud Risk Analysis (Accepted subject to minor revision due on Mar 2, 2025). Available form: https://www.sciencedirect.com/science/article/pii/S1566014125000263

[36] Wang J, Zhang S, Xiao Y, Song R. A review on graph neural network methods in financial applications. arXiv preprint arXiv:2111.15367. 2021

[37] Tavares JPHE, da Silva Medeiros ML, Barbin DF. Near-infrared techniques for fraud detection in dairy products: A review. Journal of Food Science. 2022; **87**(5):1943-1960

[38] Attigeri G, Manohara Pai MM, Pai RM, Kulkarni R. Knowledge base ontology building for fraud detection using topic modeling. Procedia Computer Science. 2018;**135**:369-376

[39] Raufi B, Finnegan C, Longo L. A comparative analysis of shap, lime, anchors, and dice for interpreting a dense neural network in credit card fraud detection. In: World Conference on Explainable Artificial Intelligence. Cham, Swizterland: Springer Nature Switzerland; 2024. pp. 365-383

[40] Yuan H, Haiyang Y, Gui S, Ji S. Explainability in graph neural networks: A taxonomic survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022;**45**(5):5782-5799

[41] Ying Z, Bourgeois D, You J, Zitnik M, Leskovec J. GNNExplainer:

Generating explanations for graph neural networks. In: Advances in Neural Information Processing Systems. New York, US: Curran Associates, Inc; 2019. pp. 9244-9255

[42] Li X, Saude J, Reddy P, Veloso M. Classifying and Understanding Financial Data Using Graph Neural Network. Washington DC, USA: AAAI Publisher; 2020

[43] Huang Q, Yamada M, Tian Y, Singh D, Yin D, Chang Y. GraphLIME: Local interpretable model explanations for graph neural networks. arXiv preprint arXiv:2001.06216. 2020

[44] Zhang Z, Liu Q, Wang H, Chengqiang L, Lee C. ProtGNN: Towards self-explaining graph neural networks. Proceedings of the AAAI Conference on Artificial Intelligence. 2022;**36**:9127-9135

[45] Yuan H, Tang J, Hu X, Ji S. XGNN: Towards model-level explanations of graph neural networks. arXiv preprint arXiv:2006.02587. 2020

[46] Xie J, Liu Y, Shen Y. Explaining dynamic graph neural networks via relevance back-propagation. arXiv preprint arXiv:2207.11175. 2022

[47] Xia W, Lai M, Shan C, Zhang Y, Dai X, Li X, et al. Explaining temporal graph models through an explorer-navigator framework. In: The Eleventh International Conference on Learning Representations. OpenReview; 2023. Available form: https://openreview.net/

[48] Huang Q, Wang X, Rao SX, Han Z, Zhang Z, He Y, et al. Benchtemp: A general benchmark for evaluating temporal graph neural networks. In: 2024 IEEE 40th International Conference on Data Engineering (ICDE). New York City, U.S.: IEEE; 2024. pp. 4044-4057

[49] Huang Q, Yan X, Wang X, Rao SX, Han Z, Fu F, et al. Retrofitting temporal graph neural networks with transformer. arXiv preprint arXiv:2409.05477. 2024

[50] Yang W, Zhang Y, Ye K, Li L, Cheng-Zhong X. FFD: A federated learning based method for credit card fraud detection. In: Big Data–BigData 2019: 8th International Congress, Held as Part of the Services Conference Federation, SCF 2019; San Diego, CA, USA; June 25–30, 2019, Proceedings. Vol. 8. Cham, Switzerland: Springer International Publishing; 2019. pp. 18-32

[51] Myalil D, Rajan MA, Apte M, Lodha S. Robust collaborative fraudulent transaction detection using federated learning. In: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA). New York City, U. S.: IEEE; 2021. pp. 373-378

[52] Ahmed A, A, Alabi O. Secure and scalable blockchain-based federated learning for cryptocurrency fraud detection: A systematic review. IEEE Access. 2024;**12**:102219-102241

[53] Peng H, Zhang Y, Sun H, Bai X, Li Y, Wang S. Domain-aware federated social bot detection with multi-relational graph neural networks. In: 2022 International Joint Conference on Neural Networks (IJCNN). New York City, U.S.: IEEE; 2022. pp. 1-8

[54] Wang Q, Pang G, Salehi M, Buntine W, Leckie C. Cross-domain graph anomaly detection via anomaly-aware contrastive alignment. Proceedings of the AAAI Conference on Artificial Intelligence. 2023;**37**:4676-4684

[55] Wu J, Xu W, Liu Q, Wu S, Wang L. Adversarial contrastive learning for evidence-aware fake news detection with graph neural networks. arXiv preprint arXiv:2210.05498. 2022

[56] Lou S, Zhang Q, Yang S, Tian Y, Tan Z, Luo M. GADY: Unsupervised anomaly detection on dynamic graphs. arXiv preprint arXiv:2310.16376. 2023

[57] Ren Y, Wang B, Zhang J, Chang Y. Adversarial active learning based heterogeneous graph neural network for fake news detection. In: 2020 IEEE International Conference on Data Mining (ICDM). New York City, U.S.: IEEE; 2020. pp. 452-461

[58] Yuan H, Zheng J, Ye Q, Qian Y, Zhang Y. Improving fake news detection with domain-adversarial and graph-attention neural network. Decision Support Systems. 2021;**151**:113633

[59] Wang L, Zhao H, Feng C, Liu W, Huang C, Santoni M, et al. Removing camouflage and revealing collusion: Leveraging gang-crime pattern in fraudster detection. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2023. pp. 5104-5115

[60] Zheng X, Wu B, Zhang AX, Li W. Improving robustness of gnn-based anomaly detection by graph adversarial training. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Paris, France: ELRA (European Language Resources Association) and ICCL (International Committee on Computational Linguistics); 2024. pp. 8902-8912

[61] Dou Y, Liu Z, Sun L, Deng Y, Peng H, Yu PS. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York, NY, USA: Association for Computing Machinery; 2020. pp. 315-324

[62] Gao Y, Wang X, He X, Liu Z, Feng H, Zhang Y. Alleviating structural distribution shift in graph anomaly detection. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. New York, NY, USA: Association for Computing Machinery; 2023. pp. 357-365

[63] Wang Z, Zhou S, Dong J, Yang C, Huang X, Zhao S. Graph anomaly detection with noisy labels by reinforcement learning. arXiv preprint arXiv:2407.05934. 2024

[64] Zhao J, Shao M, Tang H, Liu J, Du L, Wang H. RHGNN: Fake reviewer detection based on reinforced heterogeneous graph neural networks. Knowledge-Based Systems. 2023;**280**: 111029

[65] Dou Y, Ma G, Yu PS, Xie S. Robust spammer detection by Nash reinforcement learning. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery; 2020. pp. 924-933

[66] Yang R, Wang X, Jin Y, Li C, Lian J, Xie X. Reinforcement subgraph reasoning for fake news detection. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2022. pp. 2253-2262

[67] Bei Y, Zhou S, Tan Q, Hao X, Chen H, Li Z, et al. Reinforcement neighborhood selection for unsupervised graph anomaly detection. In: 2023 IEEE International Conference on Data Mining (ICDM). New York City, U.S.: IEEE; 2023. pp. 11-20

[68] Ding K, Shan X, Liu H. Towards anomaly-resistant graph neural networks via reinforcement learning. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York, NY, USA: Association for Computing Machinery; 2021. pp. 2979-2983

[69] Zhang J, Zhengjia X, Lv D, Shi Z, Shen D, Jin J, et al. Dig-in-gnn: Discriminative feature guided gnn-based fraud detector against inconsistencies in multi-relation fraud graph. Proceedings of the AAAI Conference on Artificial Intelligence. 2024;**38**:9323-9331

[70] Yang Y, Yang R, Li Y, Cui K, Yang Z, Wang Y, et al. RoSGAS: Adaptive social bot detection with reinforced self-supervised gnn architecture search. ACM Transactions on the Web. 2023; **17**(3):1-31

[71] Pateria S, Subagdja B, Tan A-h, Quek C. Hierarchical reinforcement learning: A comprehensive survey. ACM Computing Surveys (CSUR). 2021;**54**(5): 1-35

[72] Pandey M, Kaur H, Echizen I. Enhancing location privacy through prioritized experience replay in deep Q-networks. In: 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC). New York City, U.S.: IEEE; 2024. pp. 354-360

[73] Yang N, Chen S, Zhang H, Berry R. Beyond the edge: An advanced exploration of reinforcement learning for mobile edge computing, its applications, and future research trajectories. IEEE Communications Surveys & Tutorials; 2024: pp. 1-50

[74] Liang X, Chen T, Hou Z, Zhang W, Hon C, Wang X, et al. Knowledge graph-based reinforcement federated learning for Chinese question and answering. IEEE Transactions on Computational Social Systems. 2023;**11**(1):1035-1045

[75] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805. 2019

[76] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165. 2020

[77] Iyer R, Maralapalle VC, Mahesh P, Patil D. 13 generative AI and LLM: Case study in e-commerce. In: Generative AI and LLMs: Natural Language Processing and Generative Adversarial Networks. Berlin/Boston: Walter de Gruyter GmbH; 2024. p. 253

[78] Roumeliotis KI, Tselikas ND, Nasiopoulos DK. LLMs in e-commerce: A comparative analysis of gpt and llama models in product review evaluation. Natural Language Processing Journal. 2024;**6**:100056

[79] Das R, Ahmed W, Sharma K, Hardey M, Dwivedi YK, Zhang Z, et al. Towards the development of an explainable e-commerce fake review index: An attribute analytics approach. European Journal of Operational Research. 2024;**317**(2):382-400

[80] Yang L, Ott M, Goyal N, Jingfei D, Joshi M, Chen D, et al. FinBERT: A pretrained language model for financial communications. arXiv preprint arXiv: 2006.08097. 2020

[81] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;**1**(5):206-215

Section 2

Anomaly Detection in Practice

Chapter 3

Supervised Anomaly Detection with Attention

Tee Hui Teo, Chiang Liang Kok, Chee Kit Ho, Xinlong Zhang, Jovan Bowen Heng and Guangming Ren

Abstract

This chapter presents a comprehensive overview of machinery fault detection systems, focusing on anomaly detection techniques. We begin by reviewing anomaly detection and highlighting its importance in identifying irregularities in machine behavior that may indicate potential failures. The discussion then transitions to the debate between supervised and unsupervised anomaly detection methods. We analyze the advantages of supervised anomaly detection, emphasizing its ability to use labeled datasets for improved accuracy and reliability in fault identification. We further explore the prediction of remaining useful life (RUL) using advanced techniques, specifically Temporal Convolution with Attention. This approach improves predictive capabilities by allowing the model to focus on relevant temporal patterns, providing more accurate RUL estimates, and facilitating proactive maintenance strategies. Finally, we introduce an embedded machine learning fault detection system for electric fan drives. This system integrates machine learning algorithms into the hardware, enabling real-time fault detection and monitoring, enhancing the electric fan drive's operational reliability. These topics underscore the importance of advanced anomaly detection and predictive maintenance techniques in developing robust machine fault detection systems that improve operational efficiency and reduce downtime.

Keywords: anomaly detection, machine learning, statistical analysis, fraud detection, outlier detection, predictive maintenance

1. Introduction

Machines are a crucial part of modern industry, as they determine the effectiveness and safety of industrial processes. Nonetheless, machines are prone to faults caused by regular wear and tear, overloading, or adverse conditions of use. The sooner faults are detected, the better, as it can help to avoid costly downtimes and production losses, including complete machine failure. Thus, fault detection systems have become vital for industries as diverse as the high technology industry [1] and power generation [2]. In the past, equipment malfunctions have been typically resolved via reactive maintenance, meaning that the problem was addressed only when a machine broke down, and the malfunction negatively affected the machine's work [3]. While conceptually

simple, this approach was associated with lengthy downtimes and high repair costs, as easily preventable minor faults escalated into significant malfunctions. It is worth noting that according to some estimates, U.S. industries spend around $200 billion on maintaining their plant equipment and facilities, with effective maintenance-related losses of over $60 billion [4]. However, with machine learning and artificial intelligence advancements, automated fault detection systems have emerged, offering real-time monitoring, improved accuracy, and early warning capabilities [5]. This study focuses on developing an intelligent fault detection system for machines, leveraging techniques such as supervised learning and anomaly detection. By training the system with labeled datasets representing both normal and faulty states, we aim to increase the accuracy of fault predictions, reducing the need for manual interventions.

2. Review of anomaly detection

Within the data science and analytics industry, anomaly detection (also called outlier detection) identifies rare items, events, or observations that raise suspicions by differing significantly from most of the data. Anomaly detection is so important because it signals potential misconduct: anomalies can indicate incidents that could lead to security breaches, frauds, or breakages, among other things. Anomaly detection covers the convergence of multiple fields, including finance, health care, cybersecurity, and industrial automation. It also plays a central role in advancing emerging technologies such as the IoT and intelligent infrastructure management. Organizations that can quickly and accurately identify anomalies will be better positioned to act quickly to reduce risks, resulting in higher operational reliability and the potential for protecting significant assets and human lives [6].

Many different methodologies can be used to detect anomalies, each applied in a context specific to the type of data and the types of anomalies they were designed to find. These range from simple statistical tests for univariate data to complex machine-learning models in high-dimensional spaces. The selection of methods depends on the nature of the data and the use case. Several methods have been proposed and are broadly classified into classification-based, clustering-based, nearest neighbor-based, statistical, information theory-based, and spectral methods [6]. This review summarizes these approaches and highlights significant contributions in the field.

2.1 Classification-based methods

The primary focus of classification-based methods for anomaly detection is on separating normal data from the anomalies in a given data space. These approaches are typically functional under a one-class classification framework, meaning that they are designed to create a trained model that categorizes new data points as belonging to the normal class or the anomalous class. Recent advances indicate that semisupervised methods can effectively use only normal training data to detect anomalies. For example, the Deep-SVDD method uses the learned feature spaces to define what is "normal" by placing a hypersphere that encompasses the normal data points, providing an opportunity to calculate a normality score by finding the distance from the sphere built on top of this hypersphere. Additionally, self-supervised learning can improve the feature representation, with the neural networks solving extra auxiliary tasks. This can be particularly useful in situations where there is a limited amount of labeled data. Such additional tasks can involve rotations of the image or colorization of the images to create the ability

for the created features to be robust and used in different data types. Integrating these transformation-based solutions allows for expanding the spectrum of applicable data such methods can work with, including tabular datasets, and improving the applicability and performance of classification-based methods in various anomaly detection tasks [7].

2.2 Clustering-based methods

Clustering-based methods are frequently used in anomaly detection since they successfully group like data points without needing advanced knowledge about data distribution. In such methods, data are categorized based on similarity, and different distances, such as Euclidean distance, are used to create clusters. Accordingly, the significant idea behind these methods is that normal instances will result in denser clusters, while anomalies, by the definition of outliers, will remain far from these clusters. A typical example of the clustering approach is the K-means clustering algorithm, where data points are split into K separate clusters. At each iteration, the means of the members of each cluster are calculated and converge to the cluster's centroid. Accordingly, K-means can effectively group like data points. However, the choice of K significantly affects the efficiency of the cluster. In this way, hybrid approaches assume the knowledge of the decision tree classifier ID3 for anomaly detection that outperforms K-means and deals with the decision of K. Fuzzy C-means can be applied to allow membership degrees in clusters and differentiate the degrees of membership in the context of overlapping classes. Some other advanced methods are adaptive hierarchical clustering, which is intended to adapt to changing patterns of data and noise. Such examples are known to be helpful and cocluster simultaneously. It clusters data instances and clusters of parameters and offers a viable framework for the detection of anomalies [8].

2.3 Nearest neighbor-based methods

Methods based on nearest neighbors represent a time-proven approach to anomaly detection. They are designed around the idea of an occurrence of any data point being close to another point of a similar type in the feature space. In other words, all normal data instances tend to cluster regarding their features, with any unusual points outside the dense center. One of the easiest ways to evaluate such closeness is the k-nearest neighbors, or kNN algorithm, in which a test instance is examined and measured against k closest samples in training. If the average of these measurements is over the specified threshold, the point in question is labeled anomalous. Modern advances in deep learning improve the performance of these methods due to the application of a more robust feature extractor based on large ImageNet-trained datasets. These advances allow for increased performance and the creation of more discriminative features across not only image data but also other datasets, increasing the flexibility of the kNN approach. More prominently, the deep nearest neighbor anomaly detection method uses deep networks' embedding to simplify distance calculations while allowing for improved robustness in cases with limited availability of training data [9].

2.4 Statistical methods

Anomaly detection involves using statistical techniques to pinpoint data points that deviate significantly from the typical patterns found within a dataset.

Thus, these approaches must create probabilistic models for the data dynamics. Typically, they assume a specific distribution, most commonly Gaussian, to model the normal behavior of the data. Anomalies, or outliers, can then be traced as points with a low probability under the considered distribution. For univariate cases, one can use, for instance, Z-scores—data points that are beyond a certain number of standard deviations from the mean are regarded as anomalies. Or, one can use Grubb's test, which estimates the outliers iteratively, calculating the t-scores on every step—based on the sample mean and standard deviation. In multivariate cases, the Mahalanobis distance is often employed to account for correlations among variables, measuring how far a point is from the mean of the distribution [10].

2.5 Information theory-based methods

The methods based on information theory for anomaly detection are concerned with measuring the uncertainty, and information content in the dataset is used to identify anomalies. Entropy and mutual information are used to measure the distribution of data points. The other methods apply statistical measures between data points. In general, becoming an outlier is a type of high entropy behavior. For example, calculating the entropy of a random variable measures the extent of the uncertainty of the data, which is related to its structure. The Kullback–Leibler divergence is an important measure for quantifying the difference between two probability distributions. It is used to detect outliers by comparing the current distribution of data points with a reference distribution [11].

2.6 Spectral methods

Spectral methods refer to the use of mathematical information of data representations, that is, eigenvalues and eigenvectors, for detecting anomalies. Concerning hyperspectral imaging, spectral methods can be very effective as the imaging provides the spectral signatures of analyzable materials and can be used to identify anomalies on complex backgrounds. One such method is the Reed–Xiaoli method, which uses Mahalanobis distance to classify pixels with a given spectral signature as anomalies based on spectral properties [12].

3. Anomaly detection supervised or unsupervised?

There are two main types of anomaly detection methods: supervised and unsupervised. Many supervised anomaly detection methods can give a higher accuracy as they operate according to the labeled dataset where normal and anomaly instances are defined. Some methods can quickly tell normal from an anomaly [13]. Meanwhile, unsupervised data works without labeling and tries to find patterns in the data structure [14].

3.1 Supervised learning anomaly detection

Supervised anomaly detection employs labeled information to pinpoint abnormalities efficiently. This methodology explicitly characterizes both standard and unusual examples, letting the designer learn the separating qualities that differentiate these two classes amid preparation. The fundamentally preferred position of administered strategies is their ability to accomplish high location accuracy when adequate

labeled information is accessible. Notwithstanding its qualities, regulated irregularity identification is regularly confined by the requirement for generous measures of labeled information, which can be difficult and costly to acquire in numerous genuine situations. Additionally, the execution of regulated models tends to deteriorate when connected to unseen irregularities that were not spoken of in the informational collection. This impediment underscores the inherent test of adjusting these models to dynamic conditions where the nature of peculiarities may advance after some time.

Furthermore, administered anomaly identification contrasts with typical twofold order undertakings as far as the concentration is on learning the outskirts of the standard class rather than explicitly isolating two well-characterized classes. This subtle element is essential, as the irregularity class can comprise an extensive scope of beforehand obscure examples [7]. Regularly utilized strategies incorporate support vector machines, random woods, logical relapse, neural systems, and gradient boosting machines. The decision of calculation relies upon the intricacy of the information, the accessibility of labeled tests, and processing limitations [13].

Support vector machines (SVMs) are a widely used supervised learning algorithm that excels at anomaly detection tasks. SVM aims to find the optimal dividing line, known as a hyperplane, which separates normal data points from anomalous ones. Hard margin SVM represents the original formulation, working best when the data is neatly divisible into two classes without errors. In such idealized conditions, a hard margin SVM aims to draw a boundary such that all observations fall on the correct side, with none astray. However, real-world data seldom conforms to such purity, inevitably containing outliers and noise that complicate clean categorization. Under these more realistic circumstances, the inflexible hard-margin approach struggles and falters, unable to accommodate the fuzziness of imperfect information [15]. And it can be described as (Eq. (1)).

$$\begin{gathered} \min \frac{1}{2}\|\omega\|^2 \\ \text{s.t } \mathrm{y_i}\left(\omega^{\mathrm{T}}\mathrm{x_i}+\mathrm{b}\right) \geq 1, \forall_{\mathrm{i}} \\ \mathrm{h}(\mathrm{x}) = sign\left(\omega^{\mathrm{T}}\mathrm{x}\right) \end{gathered} \tag{1}$$

where is the normal vector w to the separating hyperplane, b is the bias term of the hyperplane, x_i is the training data point, y_i is the class label (+1 or − 1). Soft margin SVM introduces slack variables to address the issue that most real-world data is not perfectly separable, allowing some data points to be misclassified. While maximizing the margin between the classes, soft margin SVM also includes a penalty parameter C that controls the trade-off between maximizing the margin and minimizing classification errors. A more considerable C value reduces misclassification but may lead to overfitting, while a smaller C value increases the margin but allows more errors [15]. And it can be described as (Eq. (2)).

$$\begin{gathered} \min \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i \\ \text{s.t } \mathrm{y_i}\left(\omega^{\mathrm{T}}\mathrm{x_i}+\mathrm{b}\right) \geq 1-\xi_{\mathrm{i}}, \xi_{\mathrm{i}} \geq 0, \forall_{\mathrm{i}} \\ \mathrm{h}(\mathrm{x}) = sign\left(\omega^{\mathrm{T}}\mathrm{x}\right) \end{gathered} \tag{2}$$

3.2 Unsupervised learning anomaly detection

Unsupervised learning anomaly detection is an effective tool for data analysis that identifies unusual patterns or outliers without labeled data. By leveraging algorithms like clustering, statistical analysis, isolation forests, and autoencoders, systems can effectively uncover potential issues within large datasets. Although there are such limitations as outlier definition, scalability, and noise problems in data, the possibility of finding more opportunities and improving the quality of decision-making is considerable [14]. The K-means clustering algorithm is a partitional clustering method that divides a dataset into K clusters by minimizing the squared error between data points and their corresponding cluster centroids. For a given dataset X = xi, where i = 1,2,.. .,n and each data point xi is in d-dimensional space, the goal is to partition X into K clusters C = cj, where j = 1,2,.. .,K. The objective is to minimize the sum of squared distances by (Eq. (3)) between each data point xi and its cluster centroid μk [16].

$$J(C) = \sum_{k=1}^{k} \sum_{X_i \in C_k} \| x_i - \mu_k \|^2 \tag{3}$$

Practically, as the K-means algorithm is initialized, choosing K centroids from the dataset randomly starts. After this, each specific data point is assigned to its nearest centroid. Next the procedure of updating the centroids is developed for each cluster. They are recalculated about the mean of points. The centroids stabilize by going on this iteration, and no new changes with data points are registered regarding their cluster membership [16].

Supervised methods have achieved great accuracy when they use labeled data. However, it is limited by insufficient labeled data and cannot easily generalize the unknown anomalies in the dynamic environments. Unlike supervised classification learning, unsupervised learning (especially under incidental conditions) favors linear category structures over compact nonlinear category structures. Unsupervised learning is also multifaceted, with performance varying based on task conditions. Compared to incidental unsupervised learning, intentional unsupervised learning is more rule-like but not more accurate, and the acquisition and application of knowledge are more laborious [17].

4. Why supervised anomaly detection?

A supervised anomaly detection method involves training machine-learning models on labeled data. Such models can learn the specific characteristics of both normal and anomalous instances. The availability of labeled data mainly improves the accuracy and reliability of these models, which is why this method is highly appropriate for applications where precision is vital.

Supervised anomaly detection is discussed in the chapter, which can offer more precise modeling of anomalies compared to general approaches. By focusing on known patterns of degradation or failure, the models are better tailored to specific anomalies. Testing with additional cases or adjusting training and validation sets could provide valuable insights, but our approach aims to optimize detection accuracy for predefined anomaly types within the data. It is worth noting that the proposed model incorporates an attention mechanism, which enhances its capability to identify

and predict anomalies by focusing on relevant features. This mechanism allows the model to generalize beyond known data patterns, offering improved anomaly detection even in unseen scenarios. Thus, it provides a more robust approach to identifying significant deviations in system behaviors.

4.1 Improved accuracy in fault identification

The most essential gain of supervised anomaly detection within machine learning is its high accuracy. As the result of training on labeled datasets, the supervised models learn the specific characteristics of what constitutes an anomaly and, therefore, can effectively differentiate between normal and abnormal data. Unlike unsupervised methods, where data distribution involves some predefined assumptions, the supervised techniques provide a greater degree of precision in defining subtle anomalies. For instance, in cybersecurity, the supervised machine learning models are trained on past instances of network intrusions and, thus, can accurately identify emerging threats. At the same time, the unsupervised models may perceive some unusual but benign events as anomalies [7].

4.2 Efficient model training

Most supervised methods require less trial and error in the training phase than their unsupervised counterparts. Since the learning process is data-driven, relying on the availability of labeled samples, the model converges much more quickly on an optimal solution. As a result, business owners and IT specialists save time and computational resources. The latter benefit is precious for rapid deployment, such as real-time banking fraud detection. Moreover, the availability of a structured dataset facilitates fine-tuning the model, helping organizations effectively counteract new patterns or the evolution of the data distribution [17].

4.3 High interpretability

Being less comprehensible, unsupervised models are considered to be less transparent in comparison to supervised ones. Decision trees, linear classifiers, and other models provide some insights into how anomalies are detected. Thus, such models as gradient-boosting trees can help explain their conclusions and offer better explanations in fields such as finance and health care. For instance, fraud is one of such things: not only can decision trees trained on labeled datasets detect unusual transactions, but also they can provide why the transactions were classified as malicious [18, 19]. Thus, using supervised models can help understand some decisions to be made.

5. An embedded machine learning fault detection system for electric fan drive

This section analyzes in detail the development of the model, training performance, and real-time deployment with a focus on optimizing the CNN model for fault detection application in electric fan drives. Results are compared with a few machine learning models, and it is shown how automatic feature extraction and real-time application processing can help improve the performance of the CNN model on embedded systems.

Choosing CNN as the core algorithm offers key benefits: its lightweight architecture ensures fast response times, which is ideal for real-time applications. CNNs excel at capturing spatial hierarchies in data, making them highly effective for pattern recognition. They can also seamlessly integrate with attention mechanisms, enhancing focus on critical features.

5.1 Model development and training

This is the first stage, mainly about building up a series of machine learning models. The deep learning model CNN was trained using time series vibration data collected from a 3-axis accelerometer [20–22] under three operational conditions: Fan-Fault, Fan-On, and Fan-Off. The training was mostly a process of tuning different hyperparameters—the number of convolution layers, the dropout probabilities, and the dimensions of filters. **Figure 1** shows the split structure between training data and test data, and **Figure 2** includes CNN selected for fault detection.

5.2 CNN model performance and accuracy

After completing 30 epochs, the CNN model rapidly converged during the training process and reached a test accuracy of 99.82% and a training accuracy of 99.8%. The model achieved a good trade-off of speed and accuracy as the performance improvements beyond the 25th epoch were limited [20, 23]. This is why 30 epochs were considered a good number for finalizing the model. **Figure 3** shows accuracy over epochs, how the model's accuracy increases and loss decreases with more epochs.

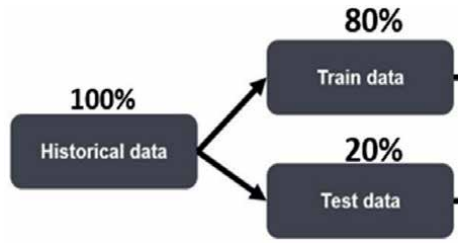

Figure 1.
Train-Test Split.

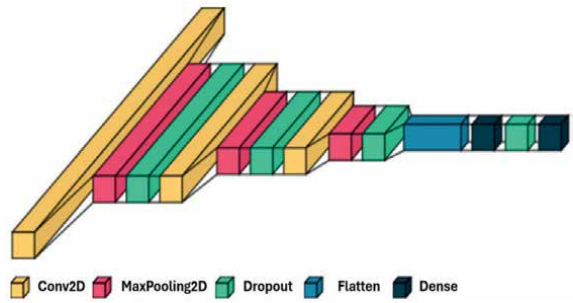

Figure 2.
Modified CNN Architecture.

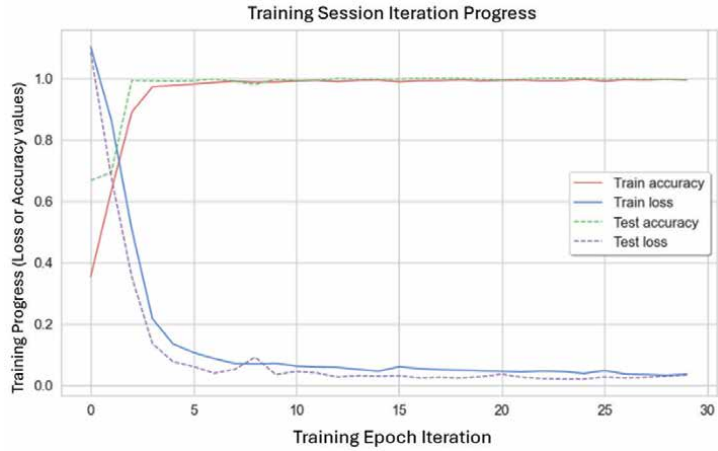

Figure 3.
Training and Validation Accuracy vs. Epoch.

5.3 Real-time testing and validation

After the model was trained, it was deployed to an embedded system on a chip, and real-time predictions were performed [24–26]. The model achieved 98, 99.8, and 99.9% accuracy on validation for the normal N-Fault, Fan-On, and Fan-Off, while it showed 100% accuracy when the case of normal data was evaluated. The system maintained high accuracy across all fan states during real-time operation (90–100%). **Figure 4** shows the precision of this model in classifying all three fan states (Fan-On, Fan-Off, and Fan-Fault) based on the entire dataset. **Figure 5** shows the classification accuracy obtained on the training set, and **Figure 6** visualizes how well the model generalizes to unseen test data. True positives and other misclassifications are visually depicted in the confusion matrices of each figure.

5.4 Comparison with traditional machine learning models

Other conventional machine learning models—SVM, RF, GB, and KNN—were also tested to confirm that the CNN performed better than them [22, 26, 27]. These were then applied to sets of raw signal data and newly derived features

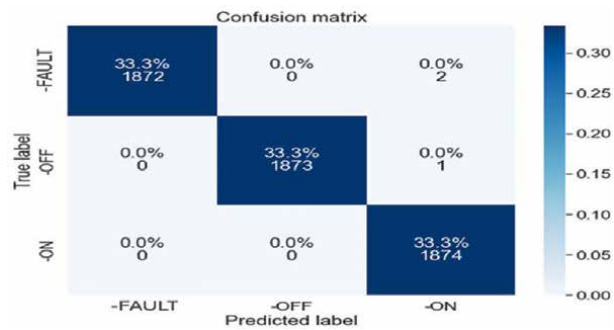

Figure 4.
Confusion matrix for full dataset.

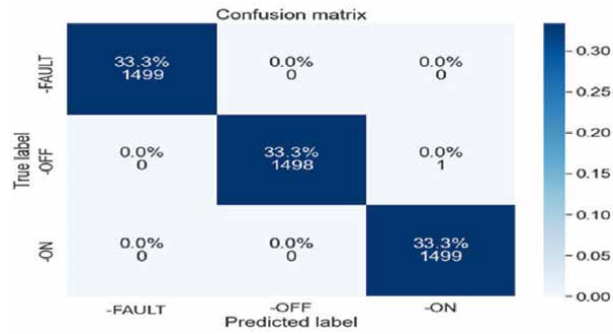

Figure 5.
Confusion matrix for train dataset.

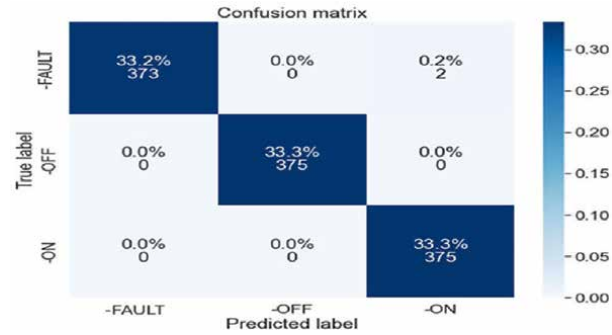

Figure 6.
Confusion matrix for test dataset.

(time-domain and frequency-domain). Despite those models providing decent performance, they were unable to achieve automatic feature extraction of the CNNs [23, 28], and they were not able to cooperate properly with raw data. Using preprocessed features, the SVM model got 94.67%, and Random Forest reached an all-time high of 96.5%, but with significant human feature engineering. To illustrate where these models go wrong, **Figure 7** compares the performance of classifiers trained on raw data.

Traditional model performance on processed features The performances of traditional models trained on preprocessed features are shown in **Figure 8**.

5.5 Performance metrics

The CNN model was evaluated with different performance metrics like Accuracy, Precision, Recall, and F1 score [25, 29]. The CNN outperformed traditional models when compared in all metrics, and it was particularly efficient in classifying the Fan-Off state with the clearest features among different classes. Fan-Fault, Fan-On, and Fan-Off F1-scores were 99.73%, 99.73%, and 100%, respectively, indicating our approach's fault detection robustness [24, 27]. While **Figure 9** shows how the CNN model performs against traditional models trained on raw data, **Figure 10** compares the CNN's performance with those of traditional models trained on statistical features.

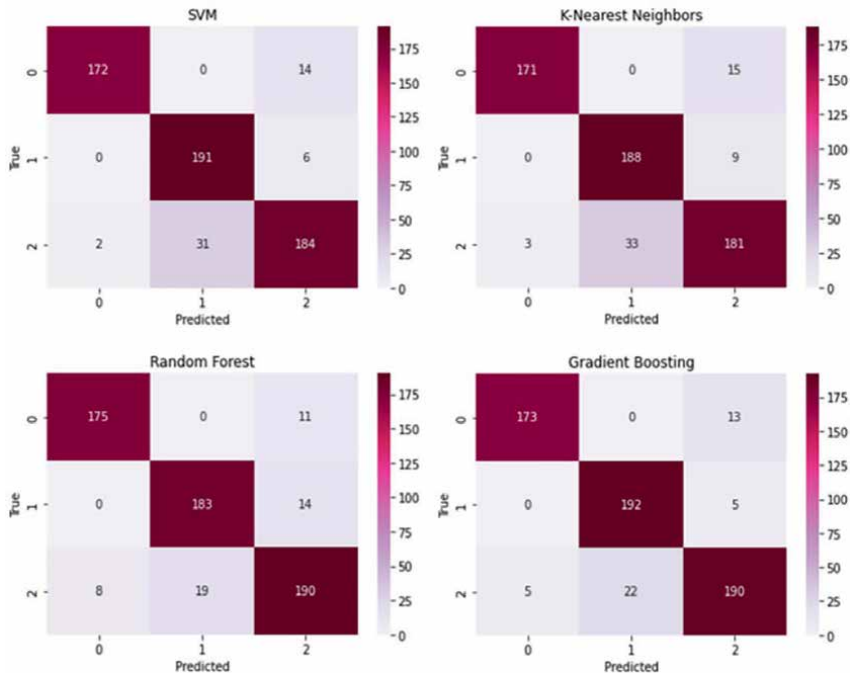

Figure 7.
Confusion matrix for first group.

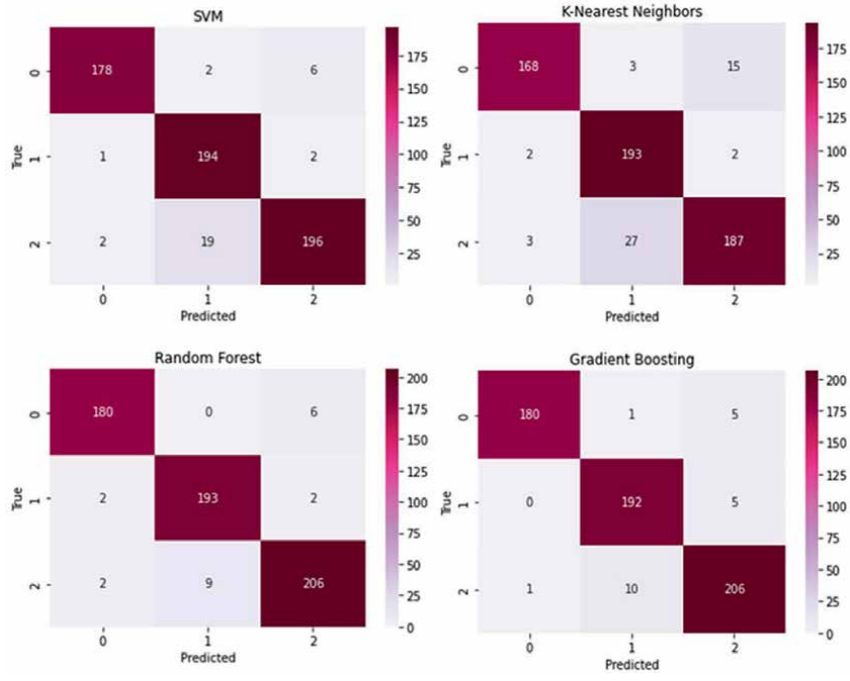

Figure 8.
Confusion matrix for second group.

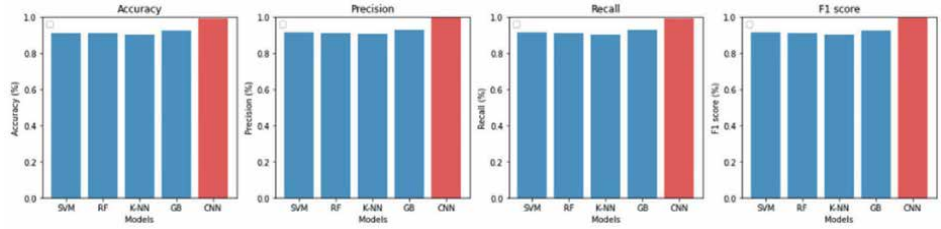

Figure 9.
Performance of CNN model vs. ML-based models (group 1).

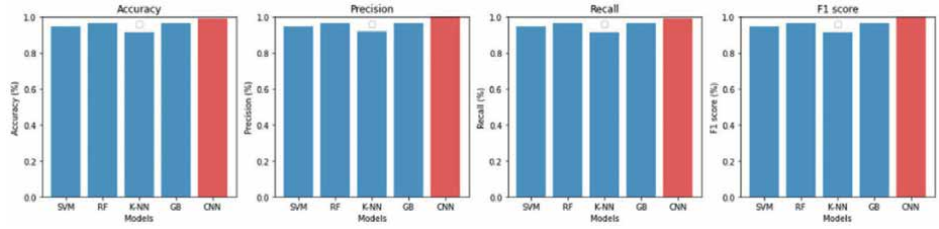

Figure 10.
Performance of CNN model vs. ML-based models (group 2).

6. Remaining useful life prediction using temporal convolution with attention

This section discusses the performance of the proposed CNN + Attention model, capturing how well it can predict the Remaining Useful Life (RUL) on the NASA C-MAPSS dataset. Most notably, the focus of our evaluation shifts from looking into computational efficiency compared to other models, and we underline the importance of attention mechanisms.

6.1 Model performance on NASA C-MAPSS dataset

The CNN + Attention model has been tested on the NASA C-MAPSS dataset, comprising four subsets (FD001, FD002, FD003, and FD004) representing different operational conditions and fault scenarios. The model was assessed for accuracy by Root Mean Square Error, Mean Absolute Error, and R-squared.

As seen in **Figure 11**, the RUL prediction results of the CNN + Attention model vs. its true values on the FD001 dataset are contrasted below. Firstly, the model produces good predictions of true values, particularly as the EOQ point is approached, which is vital in making more accurate maintenance decisions [20, 21]. On the FD001 dataset with fewer operational conditions, ADNN also achieved a lower RMSE of 10.60 than LSTM (14.28) and DCNN (12.24). Likewise, in the FD003 dataset, the CNN + Attention model reached an RMSE of 11.71, better than those of LSTM and DCNN. This shows how successful the model is in straightforward operating situations.

6.2 Proposed network structure

The architecture of the CNN + Attention model is shown in **Figure 12** below. The model consists of four one-dimensional convolutional layers responsible for processing

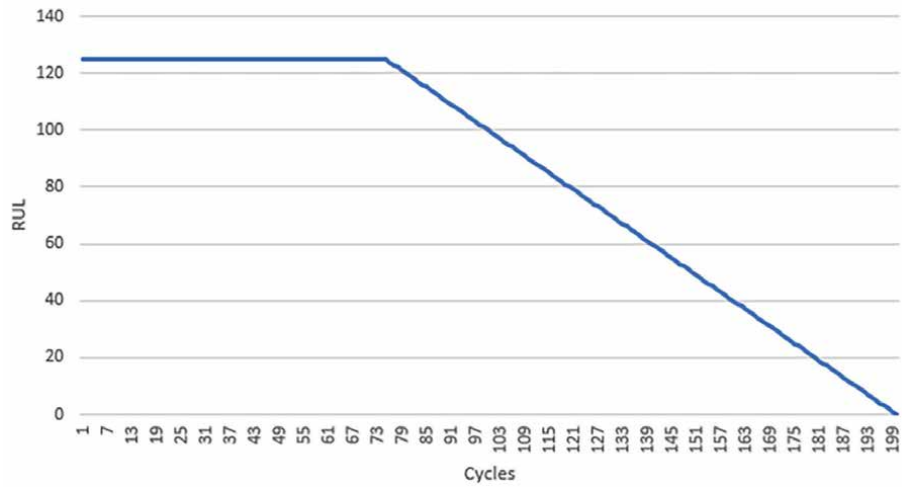

Figure 11.
CNN Layer Architecture for fault detection.

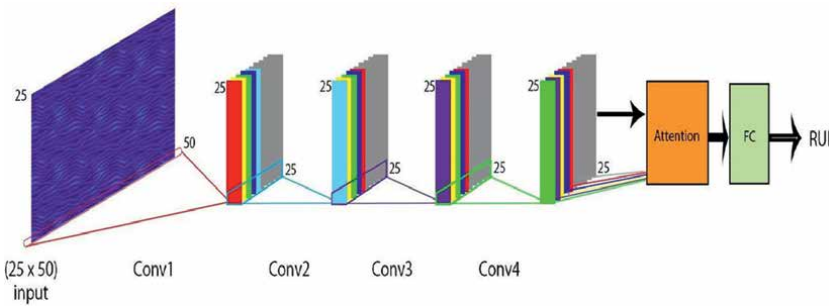

Figure 12.
CNN Layer Architecture for fault detection.

the time series data and, subsequently, an attention mechanism to capture essential features across the time domain. This leads to high-quality prediction of RUL using multivariate sensor data by making the model capture temporal elements within the data.

6.3 Model performance on complex datasets

In the case of datasets with multiple operating conditions and fault scenarios (i.e., those with complex structures, e.g., FD002 and FD004), the CNN + Attention model presented good performance compared to the other models. The training of the DBR model was finished within 45 epochs and converged to an RMSE of 14.55 for the FD002 dataset, outperforming both LSTM (18.37) and DCNN (21.02). Another case in point would be the FD004 dataset, where the RMSE was achieved as 17.23 by the CNN + Attention model, outperforming DCNN with 26.77 and being on par with LSTMs 19.65, thereby pointing to how well the model can work in a more complex surrounding.

Performance of CNN + Attention versus LSTM and DCNN across four datasets is shown in **Table 1**. It did particularly well on more straightforward datasets like FD001 and FD003 and still held its ground with competitive results for the more challenging datasets.

	NASA C-MAPSS			
Dataset	**FD001**	**FD002**	**FD003**	**FD004**
Train sets	100	260	100	249
Test sets	100	259	100	248
Operating conditions	1	6	1	6
Fault conditions	1	1	2	2
Train samples	17,731	48,819	21,820	57,522
Min/Max cycles for Train set	128/362	128/378	145/525	128/543
Min/Max cycles for Test set	31/303	21/367	38/475	19/486

Table 1.
Subsets of C-MAPSS dataset.

6.4 Computational efficiency and hardware performance

The CNN + Attention model is lightweight and can be deployed in a less resource-intensive environment. There have not been any examples of its real-time implementation, so it was implemented on a Raspberry Pi 3B to see how the model performs in real life. **Figure 13** shows the prediction times of CNN + Attention concerning LSTM as well as DCNN. Among the three models, the CNN + Attention performed best in prediction time, taking only 0.5976 milliseconds per sample compared to the LSTM (9.2114 ms) and DCNN (22.9548 ms). Due to this high computational efficiency, the proposed model can be a suitable solution for real-time prognostics, especially in industries where online monitoring of the health status of equipment is necessary.

6.5 Impact of the attention mechanism

The CNN + Attention model benefitted the most from the addition of this attention mechanism. This results in selectively focusing on the most important time steps and sensor data—with combined training, validation, and testing on C-MAPSS (all running times) being as high as 1.5 million cycles for the model to improve its

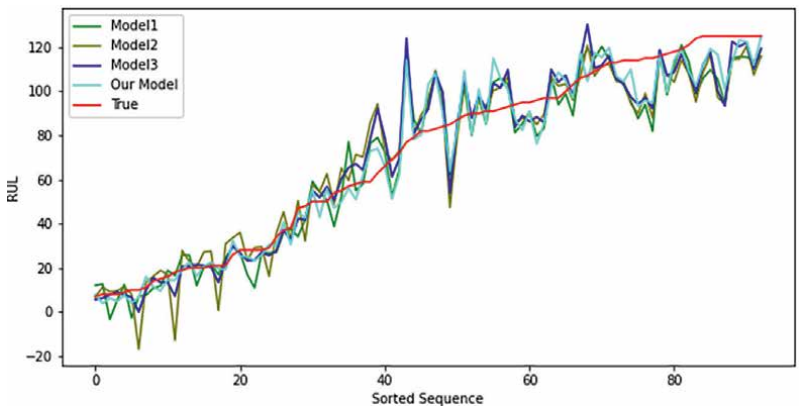

Figure 13.
Sorted True Labels of FD001 Test Data Subset (RUL versus Sorted Sequence).

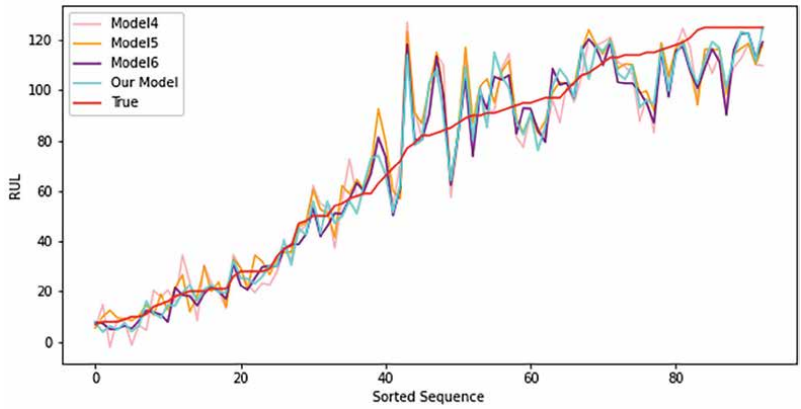

Figure 14.
Sorted True Labels of FD002 Data Subset (RUL versus Sorted Sequence).

accuracy in predicting RUL [20, 21, 26]. The improvements in RMSE and MAE with the attention mechanism as compared to without attention pattern are shown in **Figure 14**. The attention-augmented model performed significantly better.

Figures 13 and **14** plot the RUL for various models compared to the proposed model (our model) based on the data for FD001. Model1 evaluates only the convolution layers by removing the attention mechanism. Model2 changes the filter axis to convolve over time steps. Model3 reintroduces attention to Model2. Model4 applies attention to rows of the convolution output. Model5 uses softmax for attention weights instead of sigmoid. Model6 swaps sigmoid for softmax in the proposed model.

Method	RMSE FD001	Score	RMSE FD002	Score
MLP	37.56	18,000	80.03	7,800,000
SVR	20.96	1380	42.0	590,000
RVR	23.80	1500	31.30	17,400
CNN	18.45	1290	30.29	13,600
ELM	17.27	523	37.28	498,000
LSTM	16.14	338	24.49	4450
DBN	15.21	418	27.12	9030
MODBNE	15.04	334	25.05	5590
RNN	13.44	339	24.03	14,300
DCNN	12.61	274	22.36	10,400
BiLSTM	13.65	295	23.18	4130
DAG	11.96	229	20.34	2730
DSM (Regression)	14.04	310	15.15	1080
Semisupervised	12.56	231	22.73	3366
DCGAN + AE	10.71	174	19.49	2982
Proposed Model (CNN + ATT)	11.48	198	17.25	1144

Method	RMSE FD001	Score	RMSE FD002	Score
	FD003		*FD004*	
MLP	37.39	17,400	77.37	5,620,000
SVR	21.05	1600	45.35	371,000
RVR	22.37	1430	34.34	26,500
CNN	19.82	1600	29.16	7890
ELM	18.47	574	30.96	121,000
LSTM	16.18	852	28.17	5550
DBN	14.71	442	29.88	7950
MODBNE	12.51	422	28.66	6560
RNN	13.36	347	24.02	14,300
DCNN	12.64	284	23.31	12,500
BiLSTM	13.74	317	24.86	5430
DAG	12.46	535	22.43	3370
DSM (Regression)	14.62	325	21.92	2260
Semisupervised	12.10	251	22.66	2840
DCGAN + AE	11.48	273	19.71	3874
Proposed Model (CNN + ATT)	12.31	251	20.58	2072

Table 2.
Performance comparison with related work.

In addition, replacing conventional softmax with a sigmoid function for attention weight calculation allowed the model to capture multiple important features at some multiscale levels. This fix caused a performance gain on all datasets.

We compare the outputs of the CNN + Attention model using different attention mechanisms (sigmoid versus softmax) in **Table 2**, showing that our sigmoid-based approach outperforms a softmax one.

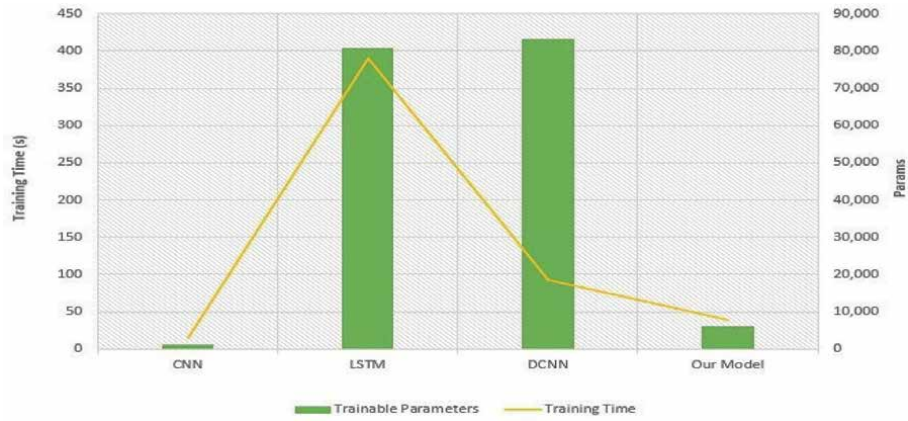

Figure 15.
Model parameters and training time for each model.

6.6 Visualizing model predictions

Figure 15 provides an additional qualitative analysis using the CNN + Attention model—a visualization of predicted versus actual RUL values for the FD001 dataset. These plots illustrate that the model can closely estimate RUL, particularly later in the engine life when it is important to make a PM decision.

7. Conclusions

The CNN model with attention mechanism proposed as a fault detection system showed better performance in traditional and complex operational environments. For example, while testing the model on NASA's C-MAPSS data in both the training and validation environment, the CNN + Attention model consistently outperformed all traditional machine learning regarding precision when detecting faults and predicting RUL. The use of attention mechanism substantially improved the characteristics of the model by enabling it to focus on critical features of the ladder, which could provide more accurate maintenance decisions in more complex operational environments, as well as in real-time applications. Furthermore, it should be noted that testing the model on resource-constrained hardware demonstrated outstandingly high computational efficiency, making it possible to use it for the real-time monitoring of faults in the production environment. Therefore, these findings suggest that using CNNs in combination with attention mechanisms can lead to the development of more reliable and easily scalable fault detection systems, which can facilitate the significant reduction of downtime and make predictive maintenance more efficient.

Author details

Tee Hui Teo[1*], Chiang Liang Kok[2], Chee Kit Ho[3], Xinlong Zhang[2], Jovan Bowen Heng[2] and Guangming Ren[2]

1 Singapore University of Technology and Design, Singapore

2 The University of Newcastle Australia, Australia

3 Singapore Institute of Technology, Singapore

*Address all correspondence to: tthui@sutd.edu.sg

References

[1] Shin HJ, Eom DH, Kim SS. One-class support vector machines—An application in machine fault detection and classification. Computers and Industrial Engineering. 2005;**48**(2):395-408. DOI: 10.1016/j.cie.2005.01.009

[2] Ali A, Khan AQ, Hussain B, Raza MT, Arif M. Fault modeling and detection in power generation, transmission and distribution systems. IET Generation, Transmission and Distribution. 2015;**9**(16):2782-2791. DOI: 10.1049/iet-gtd.2014.1023

[3] Divya D, Marath B, Santosh Kumar MB. Review of fault detection techniques for predictive maintenance. Journal of Quality in Maintenance Engineering. Jun 2022;**37**(12):14720-14728. DOI: 10.1609/aaai.v37i12.26720

[4] Amruthnath N, Gupta T. A research study on unsupervised machine learning algorithms for early fault detection in predictive maintenance. In: Proceedings of the 5th International Conference on Industrial Engineering and Applications (ICIEA). Piscataway, NJ, USA: IEEE; 2018. pp. 355-361. DOI: 10.1109/IEA.2018.8387124

[5] Zhang Y, Jiang J. Bibliographical review on reconfigurable fault-tolerant control systems. Annual Reviews in Control. Oxford, UK: Elsevier; 2008;**32**(2):229-252. DOI: 10.1016/j.arcontrol.2008.03.008

[6] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM Computing Surveys. 2009;**41**(3):1-58. DOI: 10.1145/1541880.1541882

[7] Bergman L, Hoshen Y. Classification-Based Anomaly Detection for General Data. arXiv:2005.02359. 2020. DOI: 10.48550/arXiv.2005.02359

[8] Tripathy S, Sahoo L. A survey of different methods of clustering for anomaly detection. International Journal of Scientific and Engineering Research. 2015;**6**(1):351-356

[9] Bergman L, Cohen N, Hoshen Y. Deep Nearest Neighbor Anomaly Detection. 2020;**32**(11):4824-4837 arXiv:2002.10445

[10] Srivastav A, Ray A, Gupta S. An information-theoretic measure for anomaly detection in complex dynamical systems. Mechanical Systems and Signal Processing. 2009;**23**(2):358-371. DOI: 10.1016/j.ymssp.2008.04.007

[11] Madhuri G, Usha RM. Statistical approaches to detect anomalies. In: Emerging Research in Data Engineering Systems and Computer Communications. 2020. pp. 499-509. DOI: 10.1007/978-981-15-0135-7 46

[12] Kücük F. Hybrid anomaly detection method for hyperspectral images. Signal, Image, and Video Processing. 2023;**17**(4):765-773. DOI: 10.1007/s11760-023-02492-4

[13] Bergman L, Hoshen Y, Cohen N. Selfsupervised anomaly detection in computer vision and beyond: A survey and outlook, Neural Networks. Amsterdam, Netherlands: Elsevier; 2024;**172**:106106. DOI: 10.1016/j.neunet.2024.106106

[14] Tuor A, Kaplan S, Hutchinson B, Nichols N, Robinson S. Deep Learning for Unsupervised Insider Threat Detection in Structured Cybersecurity Data Streams. arXiv.org. 2017. Available From: https://arxiv.org/abs/1710.00811

[15] Chauhan VK, Dahiya K, Sharma A. Problem formulations and solvers in linear SVM: A review. Artificial Intelligence Review. 2018;**52**(2):803-855

[16] Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, Heming J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences. 2022;**622**:178-210. DOI: 10.1016/j.ins.2022.11.139

[17] Love BC. Comparing supervised and unsupervised category learning. Psychonomic Bulletin and Review. 2002 Dec;**9**(4):829-835. DOI: 10.3758/bf03196342

[18] Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning (ICML '06). Pittsburgh, PA, USA: ACM; 2006. pp. 161-168. DOI: 10.1145/1143844.1143865

[19] Hancock J, Bauder RA, Wang H, Khoshgoftaar TM. Explainable machine learning models for Medicare fraud detection. Journal of Big Data. 2023;**10**(1):21. DOI: 10.1186/s40537-023-00821-5

[20] Wang W, Jiang H, Shao H, Wu S. An adaptive deep convolutional neural network for rolling bearing fault diagnosis. Measurement Science and Technology. 2017;**28**(9):1-10. DOI: 10.1088/1361-6501/aa6b8b

[21] Li X, Zhang W, Ding Q. Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction. Reliability Engineering and System Safety. 2019;**182**:208-218. DOI: 10.1016/j.ress.2018.10.019

[22] Zhang W, Chen D, Xiao Y, Yin H. Semi-supervised contrast learning based on multiscale attention and multitarget contrast learning for bearing fault diagnosis. IEEE Transactions on Industrial Informatics. 2023;**19**(10):10056-10068. DOI: 10.1109/TII.2022.3154168

[23] Jiang G, He H, Yan J, Xie P. Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox. IEEE Transactions on Industrial Electronics. 2019;**66**(4):3196-3207. DOI: 10.1109/TIE.2018.2864681

[24] Liu S, Jiang H, Wang Y, Zhu K. A deep feature alignment adaptation network for rolling bearing intelligent fault diagnosis. Advanced Engineering Informatics. 2022:**52**:101505. DOI: 10.1016/j.aei.2022.101505

[25] Jia L, Chow TW, Wang Y, Yuan Y. Multiscale residual attention convolutional neural network for bearing fault diagnosis. IEEE Transactions on Instrumentation and Measurement. Piscataway, NJ, USA: IEEE; 2022:**71**. DOI: 10.1109/TIM.2022.3155057

[26] Bai J, Ding B, Xiao Z, Jiao L. Hyperspectral image classification based on deep attention graph convolutional network. IEEE Transactions on Geoscience and Remote Sensing. Piscataway, NJ, USA: IEEE; 2022:**60**. DOI: 10.1109/TGRS.2022.3149672

[27] Cheng M, Lin J, Lu S, Dong S. Seismic data reconstruction based on multiscale attention deep learning. IEEE Transactions on Geoscience and Remote Sensing. 2022;**60**. DOI: 10.1109/TGRS.2022.3194729

[28] Zhao K, Jia F, Shao H. Unbalanced fault diagnosis of rolling bearings using transfer adaptive boosting with squeeze-and-excitation attention convolutional

neural network. Measurement Science and Technology. 2023;**34**(4):045107. DOI: 10.1088/1361-6501/ab4569

[29] Galassi A, Lippi M, Torroni P. Attention in natural language processing. IEEE Transactions on Neural Networks and Learning Systems. 2021;**32**(10):4291-4308. DOI: 10.1109/TNNLS.2021.3062893

Chapter 4

Anomaly Detection in Metal-Textile Industries

Ingo Elsen, Alexander Ferrein and Stefan Schiffer

Abstract

In this paper, we presented an approach to deploying a student–teacher feature pyramid model (STFPM) for anomaly detection metal-textile gas filters used in automotive exhaust gas filtering at GKD-Gebr. Kufferath AG. As the customer requires 100% quality of the delivered parts, an optical inspection process of every produced filter is required. This is very demanding for the human inspection worker as she has to inspect many 100 parts in an 8 hours shift. On the other hand, a fully vision-based system is not able to achieve the required classification rates either. Therefore, we propose a one-class anomaly detection process for the gas filters where human and AI work together in achieving the 100% pass rate. The STFPM model deals with the large amount of clearly true positive cases and automatedly classified them as PASS. Only cases of doubt where an anomaly has been detected are inspected by the human inspector. This way, the work load of the inspection worker is reduced, and, on the other hand, the hard to meet case of no mis-classification of the AI system can be avoided. We show the network architecture and the integration into the quality inspection process of the company GKD.

Keywords: artificial intelligence, computer vision, machine learning, quality control, process optimization

1. Introduction

The use of image processing techniques for industrial quality inspection systems has a long tradition and is applied in the geometric measurements of produced parts as well as defect detection, contamination detection, and feature adherence either in 2D and 3D [1–3]. With the availability of computing power and ever increasing storage capacity, the application of machine learning approaches in this field has become feasible for use cases where classical industrial image processing failed. One of these areas is metal textiles that pose multiple challenges on the image processing pipelines when it comes to the quality inspection of the textiles and industrial parts than integrate these textiles, respectively.

Machine learning approaches would formulate this quality inspection task as a classification problem. Typically, the machine learning model is trained on a dataset that contains images of the respective classes in a more or less even distribution. However, many times this conflicts with the production process itself: Firstly, the

datasets that can be generated will almost always be highly skewed to the class of parts that are error free (called PASS parts, and FAIL for the faulty parts). This is obvious, as otherwise the production would not be economically attractive. Secondly, the FAIL parts often must be decomposed into multiple classes themselves, e.g., weave errors, dents, welding errors, etc. This increases the class distribution imbalance with respect to the FAIL class even more. For all FAIL classes, further decisions and actions have to be taken, e.g., scrapping vs. repairing on a per class level. These decisions and actions often involve the assessment of the degree of damage in the FAIL parts and the existing processes for further treatment of parts belonging to these classes. This knowledge is often available only implicitly in the experience of the quality control inspectors. Hence, instead of trying to address this multi-class problem with a fully automated solution a one-class problem, using anomaly detection in a human-in-the-loop process is more promising.

As a main contribution of this paper, we describe a use case of an optical inspection process in metal-textile industries which need to provide a nearly 100% PASS rate. We propose an approach where the automated inspection system reliably detects parts that are undoubtedly PASS parts and only sends cases in doubts to the human quality inspector. This way, the inspection payload of the human inspector, i.e., the number of parts to be inspected by the human inspector, is reduced. This in turn leads to a better quality of delivered parts and better working conditions for the inspection worker. This work is an extended version of our previous work published in [4].

The rest of the paper is organized as follows. In the next section, we review some related work with respect to optical inspection and anomaly detection. Section 3 presents our solution proposing a STFPM architecture for a one-class anomaly detector and shows how it has been integrated into the inspection process of the GKD company. In Sections 4 and 5, we discuss our work and conclude with an outlook on future research and future application areas.

2. Related work

In the following, we review work that is related to our work from several different perspectives. Our approach aims at reducing the worker's load in the process of quality inspections while we remain at a quality level of near to 100%. Therefore, we review related work showing other applications of the Six Sigma approach taken in our metal-textile application use-case first. Then, we review works from the area of making decisions in optical inspections systems, and in general, an review works from the area of anomaly detection systems, in particular.

Six Sigma is a set of techniques and tools for process improvement introduced for manufacturing training programmes in the late 1980. While its original focus was on process electronics industries, it spread out to many sectors and is still widely used today. It comes with a set of statistical tools for improving existing processes. Statistical tools help identifying non-conforming products in the range of parts per million (PPM). It means, on the other hand, that establishing a process following the Six Sigma approach, the output of defective parts is at the level of 3.4 PPM and below [5]. Six Sigma approaches in production settings involve a quality control process that is integrated in the production process [6]. Quality control can be performed automatically, manually or as a mixture of both approaches. If a manual inspection is part of the quality control process, the process itself loses its stationary nature due to human factors induced, such as stress, fatigue, or individual variances. These have their cause

in the repetitiveness of the process and the required long time spans of concentration required by the worker. It turns out that manual quality control can only detect between 60 and 80% of defective parts. At the same time, the quality inspection process takes up to 10% of the whole labour costs [7]. The above-mentioned factors can even result in a performance reduction below this value [8]. Furthermore, the structure of this work can have negative impact on the workers' health, especially when visual inspection is involved [9]. With such detection rates, it is hard to establish a Six Sigma process based on human inspection alone. In a fully automated process, on the other hand, this factor depends on the performance of the underlying algorithm, which can also not achieve the required quality of a Six Sigma quality control process. Therefore, a combined approach as proposed in this work seems beneficial to come closer to the goals of achieving a Six Sigma quality control process. Usually, the final stage in a typical Six Sigma process includes the inspection of the final product to check either its quality and/or the process [7]. In this paper, we focus on the control process which, in turn, focuses on a single product entity checked by optical inspection.

Statistically, an optical inspection system (OIS) for quality control can be seen as a classifier with a specific operating characteristic. The performance goal of the quality control system is to distinguish passed from failed parts while reducing the total number of wrong classifications. The amount of admissible errors depends on the application domain. Six Sigma allows for 3.4 PPM of false negative decisions (part is FAIL, but decision is PASS) (cf. the red area under the gray curve in **Figure 1**). The number of false positives (part is PASS but decision is FAIL) should be as small as possible to reduce the negative economical impact. An OIS relies purely on visual inspection of the parts produced to make this decision. This can either be a simple image taken by a camera but can, such as in our case, also be a combination of different types of digital images and a visual inspection by a human quality control expert.

Automated OIS can be distinguished into parameter-based systems that check specific parameters such as dimensions based on a static definition of parameters. These are typically hard coded into the inspection algorithm and require a calibrated camera system. Model-based systems check a part against an expected visual

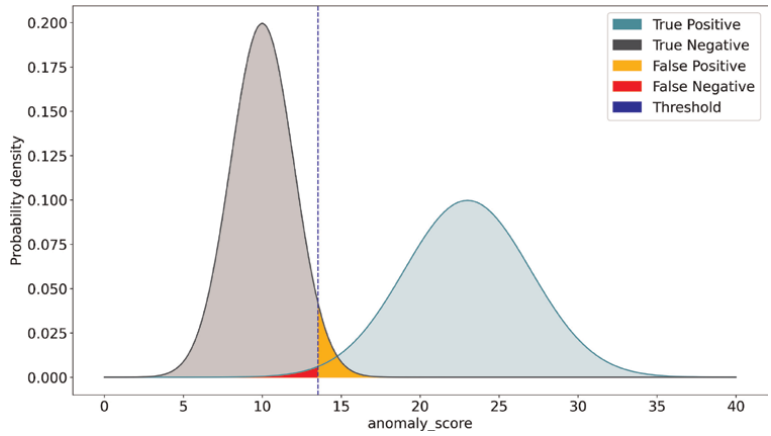

Figure 1.
PASS (left of threshold line) and FAIL (right of threshold line) decision process based on thresholding. The false negatives are critical, as these must not violate the Six Sigma conditions of 6 ppm.

appearance for both possible decisions. The decision takes the most probable of the two classes or checks against the PASS quality level by thresholding the deviation from the optimal case. This approach is known as *anomaly detection* (AD).

The approaches of anomaly detection in OIS can be separated in three different areas: (1) one-class classification, (2) reconstruction-based methods, and (3) representation-based methods. In a typical production process, a high amount of normal products are produced. This is a challenging task for modeling anomaly detection, because it generates a highly imbalanced dataset with much more images of PASS parts than images of FAIL parts. Most state-of-the-art solutions are working with convolutional neural networks (CNN) that require large training sets (see, for instance, [10]). To make a machine learning classifier unbiased in its decision, the datasets should be balanced with respect the number of examples in the different classes. Alternatively, the classification problem can be reformulated as a *one-class-classification* problem that is trained only on the normal (PASS) images that count for the majority of the data in the dataset. The most common methods for one-class classification are, for instance, shown in [10–12]. For training the network, high quality and in the best case balanced datasets are crucial and not easy to get. In [13], a dataset of about 1000 images of aluminum part with 102 defect-free and 23 defective images is presented. Rippel and Merhof [14] mentions (besides giving an overview of the field of anomaly detection methods) four more publicly available datasets for anomaly detection with a dataset size ranging from 1300 to 5300 images. In the area of detecting anomalies for photovoltaic cells, [15] proposed a large dataset which consists of more than 36.000 images of good and defective photovoltaic cells with eight different defective categories.

The main two classes of CNN-based anomaly detection methods are *reconstruction-based methods* and *representation-based methods*, both with capabilities for image-level and pixel-level predictions. In the former case, we speak of anomaly detection while in the pixel-level case one speaks of anomaly segmentation. This is based on the assumption that a model which was trained with images of O.K. parts is not able to reconstruct defective parts as well. In representation-based methods, anomalies are being detected on distributions from the extracted features of a neural network. Reconstruction-based methods, on the other hand, are founded on generative models. The three main models are auto-encoder (AE), generative adversarial network (GAN), and normalizing flow networks (NF), where NFs can be seen as a mixed form. In traditional AE, the input and the output of the network are compared with L2 norm or structural similarity loss. Other authors who deployed GANs used generated images for training. The high generalization capabilities of GANs are restricted introducing pseudoanomalies in a self-supervised fashion [16, 17]. Especially, self-supervised methods shift the problem of the high generalization to a bias of pseudoanomalies, which results in poor performance in benchmarks.

A special case of reconstruction-based models are normalizing flow models, which show very good results and have the capabilities to estimate likelihoods by learning transformations between densities and given distributions. The learning is done by an invertible mapping function which transforms basic probability functions in multiple steps into the target distribution. By being invertible, data sampled from this learnt representation can be projected back into the original space. Representation-based methods differ from reconstruction-based methods in their comparison between normal and abnormal images which is done in the feature space instead of the image space. They typically have two parts: feature extraction and feature comparison between data points and expected distributions. Because of the separation between

feature extraction and comparison, there is a big freedom in choosing the neural network backbone. Rippel et al. [18] showed in their work that neural networks which were pre-trained on ImageNet can indeed generate meaningful features. Pre-trained features of single layers in a pyramidal way were proposed by Cohen and Hoshen [19]. The training is done by storing all layer-wise aggregated features. For inference, a simple kNN search is done which yields a patchwise, multiscale feature comparison with the maximum distance as anomaly score. Aggregating the ideas of Cohen et al. and Rippel et al. [18] by simply calculating a Gaussian on every patch and taking the Mahalanobis distance between inferred and stored features, Defard et al. [20] improved the performance on the MvtecData set in PaDIM by reducing kNN search time in combination with a randomized dimension reduction. Due to the Gaussian-based outlier sensitivity, PatchCore [21] introduced core set sampling which reduces the kNN search space up to 99%. This family of algorithms has been further improved: SOMAD [22] puts a self-organizing feature map in the PaDIM setup, SA-PatchCore [23] is using transformer-based self-attention mechanisms to get a more global receptive field of view for detecting co-occurence anomalies, and Kim et al. [24] sped up the calculation of covariance matrix in PaDIM. Another group of methods is the group of knowledge distillation-based approaches. In these approaches, an untrained student network is trained to learn the layer-wise feature representations of a pre-trained teacher network, which has often the same architecture as the student network and is shown only normal images. The assumption is similar to the one in reconstruction-based networks: By minimizing the difference in the representation on only normal images, the difference between student and teacher should be larger when an abnormal image is shown. An important method which we also deploy in our work presented here is the student-teacher feature pyramide matching (STFPM) proposed by Wang et al. [25].

3. Technical solution

In this section, we describe our technical solution for the human-AI coworking optical quality inspection process for metal-textile filters by GKD-Gebr. Kufferath AG. In the next section, we first recap prerequisites that need to be met for a data-driven project following a CRISP-DM approach. In Section 3.2, we outline the current process and how the data for training our STFPM network were acquired. In Section 3.3, we define our model and show how it is being trained. Then, the integration of our approach into the QC process of GKD is explained in Section 3.4.

3.1 Project requirements following CRISP-DM

According to the standard process for this type of data driven projects, CRISP-DM [26], the requirements for a technical solution must be defined during the *business analysis* process step. These requirements must be quantifiable, if possible, and include also the requirements for the potential deployment, i.e., the operation of the designed solution. The core requirements and constraints are listed in very condensed form in **Table 1**.

As has been discussed in [27], the use of any particular method from artificial intelligence imposes a set of specific requirements. In our case, rather generally speaking, all (supervised) machine learning methods need data to be trained. This is especially true for image-based classification tasks like in our case here. What is more,

Name	Rationale	Quantification
Reduction of workload	QA workers shall be relieved from analyzing filters that are safely PASS parts, where safely means, that the Six Sigma conditions are not violated	Reduction: **Min:** >0% **Plan:** >50% **Wish:** >66%
No degradation in 'fail' rate	Even if Six Sigma conditions are kept, the rate of 'fail' parts should not increase	Fail rate: **Min:** <5% **Plan:** <4% **Wish:** <3%
Reduction of false positives	There is an assumption that at the end of shifts, the fatigue of QA workers leads to unwanted 'fail' decisions (false positives). With reduced workload, there should be an improvement in the human decision process.	FP reduction: **Min:** <0% **Plan:** <1% **Wish:** <2%
Application in production process	The integration of the anomaly detector should not slow down the overall production process, even if labelling the human-controlled parts is included	Rate reduction: **Min:** <5% **Plan:** <1% **Wish:** <0%
Minimize setup modifications	Modifications to the existing production environment should be kept as small as possible	Modifications: **Min:** Mechanics, Touchscreens, network image storage **Plan:** N/A **Wish:** N/A

Table 1.
Requirements for the use case. Setup modifications were fixed during analysis, and hence, there is no variation in potential execution.

Figure 2.
Example for a recirculation filter. Besides defects in the mesh, defects induced during welding of the ring might cause defects. Source: GKD-Gebr. Kufferath AG.

the quality of the training data drastically influences the quality of the resulting model and in turn the quality of the solution and the success of the overall application (**Figure 2**).

3.2 Current process and dataset generation

The technical solution for the use cases starts with the generation of a dataset that can be used for training an image-based anomaly detector. Based on a previous failed attempt to inspect the parts automatically with classical industrial image processing approaches, a setup that could capture images of the parts produced already existed in the company (cf. **Figure 3**).

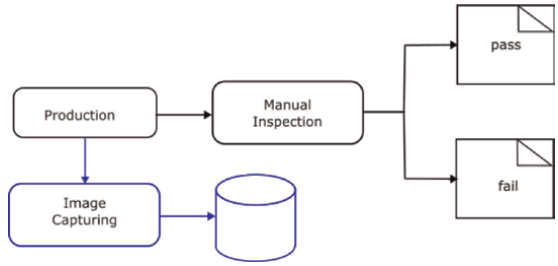

Figure 3.
Current production process. 100% of parts are manually inspected. PASS and FAIL decisions are completely human driven. The data flow is marked blue.

The filters are produced in a closed manufacturing cell where two line cameras are capturing the filter while it is under production on a turntable. The first camera captures the gray values, and the second camera captures the distance of the filter elements to the camera's chip plane. Thus, two images showing the unwind filter are available per produced part with 10,500 × 1408 pixels (**Figure 4a** and **b**) and 10,500 × 1409 pixels for the depth image, respectively (**Figure 4c** and **d**). Images are horizontally and vertically aligned, so that there is a direct correspondence between image regions.

While the image capturing happens in-line with the production, the speed of production is determined by the speed of inspection by the QC-workers that have to manually inspect 100% of the filters produced.

Production errors can occur in

- the mesh,
- the transition between mesh and weld,
- the weld,
- the transition between weld and shell, and
- the shell.

Hence, errors are possible in image parts that are highly textured and low structured and vice versa.

3.3 Model building for the anomaly detector

While there are well-established approaches for anomaly detection, e.g., local outlier factor [28], one-class support vector machines [29], and isolation forest [30], to name a few, an anomaly detector for image-based human-in-the-loop (HITL) approaches should not only take a decision if an anomaly exists. It should, if possible, also show the position of the anomaly detected to guide the quality control worker in the further inspection process.

The latter requirement can be fulfilled by reconstruction-based methods and knowledge distillation-based method, as mentioned in the related work section. The well-known texture affinity (over structure) of convolutional neural networks [31] is beneficial for the problem at hand. While the woven metal mesh can be seen as a large

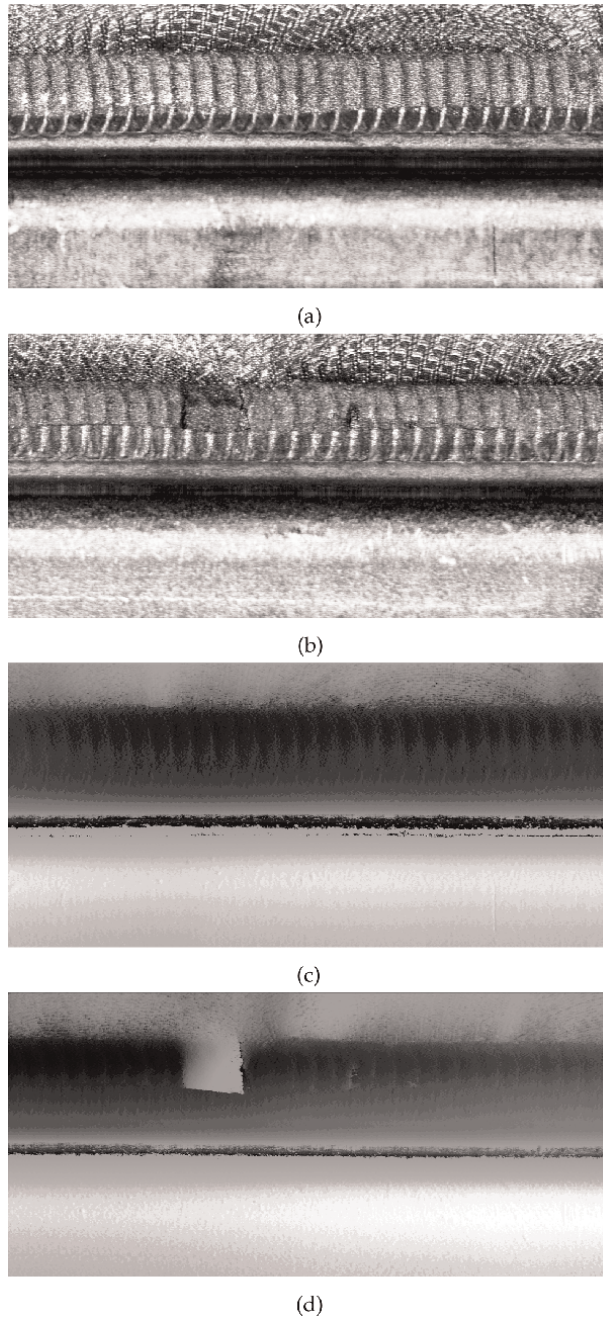

Figure 4.
Examples for non-defective and defective filters images. Source: GKD-Gebr. Kufferath AG. (a) Monochrome image of non-defective filter, (b) monochrome image of defective filter, (c) depth image of non-defective filter and (d) depth image of defective filter.

texture, the welded areas of the filter have virtually no texture at all, which hold true for the depth image as well.

The machine learning model used here is a student–teacher feature pyramid model [25] (STFPM) that gets trained on images of PASS class filters only. Alternatives have

also been researched and are subject for continuous improvement [4]. The local reconstruction error can be visualized using heat map colormaps. This reconstruction error image is shown to the quality control workers to guide them to the potential source of the anomaly. As shown in **Figure 5**, the anomaly is exactly at the predicted position.

3.3.1 Training

The STFPM training uses two networks with an identical architecture. Each network receives the input image, or a batch of input images, respectively. The training set consists of PASS images only. A preprocessing stage performs a patching of the images to reduce the input size and the total size of the networks parameters. The teacher network (marked blue in **Figure 5** is pre-trained with frozen weights, while the student (marked orange in **Figure 5**) has its weights randomly initialized and gets trained. The target of the student is to reproduce the output in the latent spaces by minimization of error on the different scales (l^1, l^2, l^3).

3.3.2 Inference

During inference, both networks receive the input image. Now only the component-wise differences are taken. At the respective latent layers (l^1, l^2, l^3), the differences can be seen in **Figure 5**. Each feature map is scaled up to the size of the lowest feature map (l^1) $(\uparrow)$, and the results are component-wise multiplied (Π), producing the output image as map of the reconstruction error.

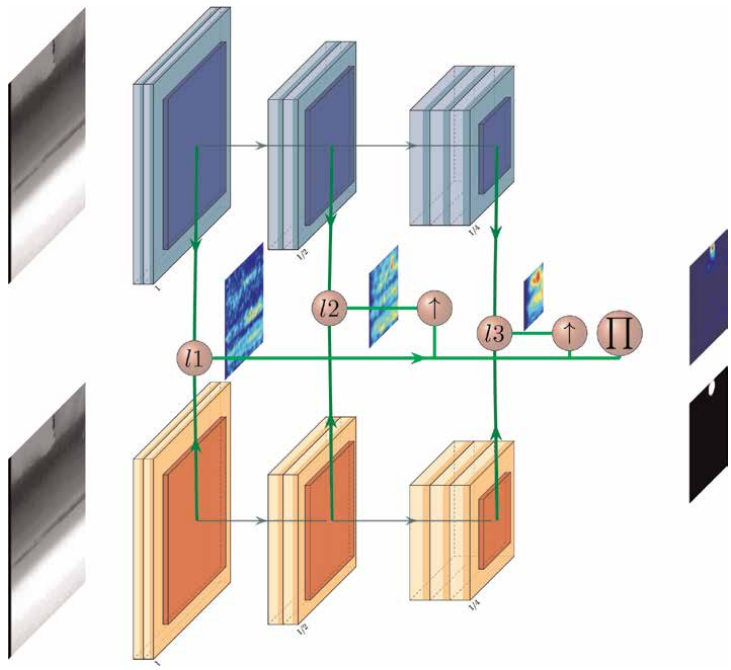

Figure 5.
STFPM approach to anomaly detection. The reconstruction error heat map indicates the position of an anomaly and is presented to the worker.

Figure 6 gives a more detailed view on the anomaly heatmap. While the welding error is easily to spot for humans, the small mesh error only becomes apparent in the heat map image.

3.4 Process integration

Process integration may not only consider the production process itself, but also the influence of the process change to the human personnel involved in the total process. To achieve that, a development process was designed that integrates the technical part, together with the organizational and process part plus the analysis of human factors [32]. This is necessary as a human-in-the-loop approach must consider the mutual influences of the human and technical parts of a solution. Using this method, the final process was designed and changes to the existing process steps applied. The final process is shown in **Figure 7**.

The anomaly detection process now involves a predefined decision boundary for the anomaly score. This boundary determines whether a part is identified as PASS or for further inspection 'INSP' (**Figure 7a** and **b**). As can bee seen from **Figure 7**, the newly introduced class 'INSP' includes all parts, where the anomaly score is exceeding the threshold. This still includes a subset of correctly produced parts, which would lead to an Type 1 classification error in a fully automated system. For these parts, the QC worker has the final call and can also take further action depending on the class and severity of the error. This reduces scrap and hence improves the process' efficiency.

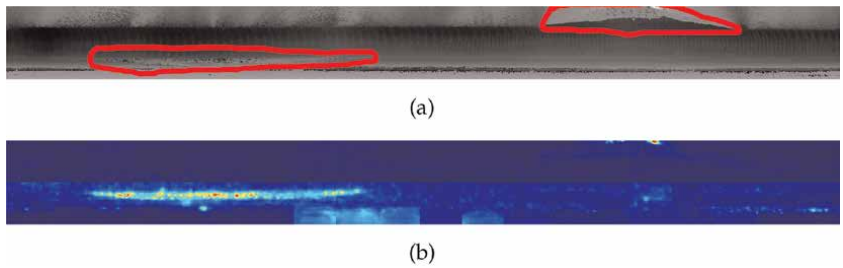

Figure 6.
Image with multiple errors and the calculated activity map. (a) Depth image with anomalies: red area (manually drawn for illustrative purposes) is showing the area of the anomalies and (b) anomaly map generated with the semantically tiled STFPM model.

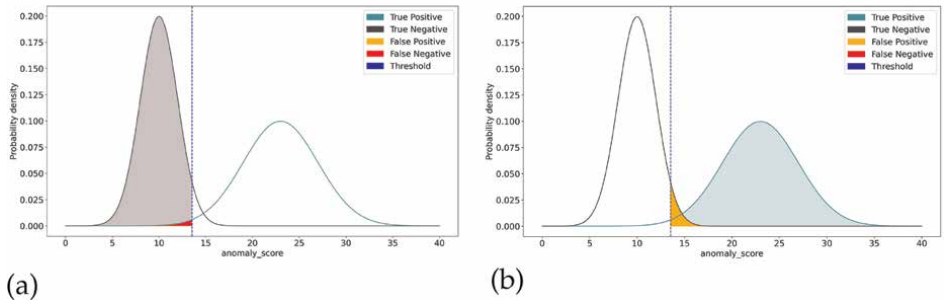

Figure 7.
Automated decision process to reduce worker load. (a) Parts identified by ML model as passed and (b) parts isentified by ML model as "for further inspection".

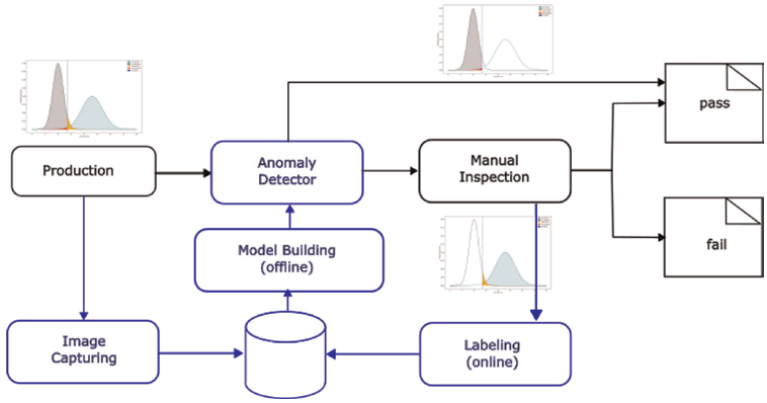

Figure 8.
HITL approach using anomaly detection (AD) and human decision process of non-pass parts. Image labeling is used for later stages to replace AD with a classifier model.

The PASS parts that come out of the anomaly detection stage contain a portion of type 2 classification errors. These must not exceed the Six Sigma threshold including a safety margin for errors that might be introduced by the QC worker.

The whole HITL process can be seen in **Figure 8**. As depicted, the part of the reduction in workload for the QC worked can be used to label the data of the manual inspection. Thus, over time, the dataset is augmented with properly labeled data. This labeled data, although small in size, can be used for automatic threshold determination for the anomaly detector. So instead of designing the anomaly detector with a safe margin threshold, this threshold can be constantly rebuilt, taking the labeled dataset into account. This reduces the workload even more while keeping the quality requirements fulfilled.

4. Discussion

As it turn out, quality control processes following a Six Sigma approach, where the optical quality inspection has to sort out all (or near to all) defective parts, are hard to come by. Human inspection can only reach up to 60–80% inspection rate. Today, even with a technical AI-based solution, the margins of Six Sigma are very hard to meet. We therefore propose a combined approach where the optical inspection is done by the human only for parts that surely are non-O.K. parts. Instead of inspecting 100% of all parts, only a fraction of parts need to be inspected by the human worker following the STFPM approach proposed in this work. This way, the cognitive load of the worker is reduced during her shift alleviating the process to find all of the defective parts that have been produced in her shift. While definitely the stress level for the inspection worker is reduced, it is still to be shown to which level the worker is relieved. This will be further investigated in future works. As for now, we can only state that the QC worker involved in this work was positive about the reduction of the number parts to be inspected. While we get positive feedback from the shop floor, there are some challenges that need to be further addressed in the future. One of them is the problem of interfering with a running inspection system. The worker are used to their processes and need to instantly switch to a different process. Another complication is that the visual representation of the anomaly from the technical inspection process is not familiar to the worker in the first place. Here, we need to find good and intuitive representations for the worker in the QC process.

Another important issue is to also look into other possible models such as ROCKET [33] which make use of random convolutional kernels for time series classification. First experiments of deploying such an architecture for 2D data look very promising.

5. Conclusion

In this paper, we presented an approach to deploying a student–teacher feature pyramid model for anomaly detection metal-textile filters used in automotive exhaust gas filtering. The quality requirements are following a Six Sigma approach, i.e., no defect parts may be shipped to the customer. The quality inspection was done 100% visually by a human inspection worker as technial inspection systems installed at the production plant could not meet the quality requirements by themselves. We therefore redesigned the QC process in such a way: an STFPM-based anomaly detector trained on good and defective parts is separating the surely PASS parts from other parts that might have problems, i.e., where an anomaly has been detected. Only those parts are further inspected by the human inspector. First tests show that an reduction of workload of 50% is achievable, with a margin left to further improvement without violating the Six Sigma requirements. These values have been measured in multiple testing scenarios with a smaller dataset of 352 test images. The threshold for the anomaly decision was then moved so that no false negatives were produced. The true positives (pass parts) were then upscaled to the real production values, from which a workload reduction of 69.4% was calculated. Even when an additional safety margin is added, reduction was above 50%. Final results will be taken on a larger set of images from a longer period of production that will also account for parameter shifts, e.g., in the cameras and lighting.

In conclusion, we see a very positive way forward of deploying AI systems on the shop floor by following a human-in-the-loop approach. The AI-based inspection system deals with the large mass of undoubtedly PASS parts and detects any form of anomaly. With the human in the loop, there is no need to further specify different fault classes, and a one-class classification process is sufficient. This has also the positive effect that the training process of the AI-based anomaly detector is simplified as fewer data for distinguishing different anomaly classes is required. This also reduces the development costs of such systems for industry.

In this paper, we presented an extended version of the metal-textile anomaly detection use case which was presented in [4]. While we are very convinced that the approach is a way to help automating the QC process on the shop-floor, we need to further investigate on our work. For instance, we need to survey how much the automated inspection process relieves the worker and how this influences the outcomes of the quality process. Further as already mentioned, we also will look into other promising network architectures in our future work.

Acknowledgements

We acknowledge the support by the Federal Ministry of Education and Research (BMBF) under grant no 02L19C602 and GKD-Gebr. Kufferath AG for their cooperation and permission to use their images. We would like express our gratitude to our past and current WRIKsam staff for their contributions in this work, in particular, we would like to thank T. Arndt, M. Conzen, O. Galla, H. Köse, and M. Tschesche.

DOI: http://dx.doi.org/10.5772/intechopen.1008411

Author details

Ingo Elsen, Alexander Ferrein* and Stefan Schiffer
Mobile Autonomous Systems & Cognitive Robotics Institute, FH Aachen—University of Applied Sciences, Aachen, Germany

*Address all correspondence to: ferrein@fh-aachen.de

IntechOpen

References

[1] Demant C, Garnica C, Streicher-Abel B. Industrial Image Processing. Berlin, Heidelberg, Germany: Springer; 2013

[2] Fabijanska A, Kuzanski M, Sankowski D, Jackowska-Strumillo L. Application of image processing and analysis in selected industrial computer vision systems. In: 2008 International Conference on Perspective Technologies and Methods in MEMS Design. New Jersey, NJ, USA: IEEE; 2008. pp. 27-31

[3] Aguilar J-J, Torres F, Lope M. Stereo vision for 3d measurement: Accuracy analysis, calibration and industrial applications. Measurement. 1996;**18**(4): 193-200

[4] Arndt T, Conzen M, Elsen I, Ferrein A, Schiffer S, Galla O, et al. Anomaly detection in the metal-textile industry for the reduction of the cognitive load of quality control workers. In: Proceedings of the 16th International Conference on Pervasive Technologies Related to Assistive Environments. New York, NY, USA: Association for Computing Machinery/ACM; 2023. DOI: 10.1145/3594806.3596558

[5] Tjahjono B, Ball P, Vitanov V, Scorzafave C, Nogueira J, Calleja J, et al. Six sigma: A literature review. International Journal of Lean Six Sigma. 2010;**1**(3):216-233

[6] Blanco-Encomienda FJ, Rosillo-Díaz E, Muñoz-Rosas JF. Importance of quality control implementation in the production process of a company. European Journal of Economics and Business Studies. 2018;**10**(1):248

[7] Newman TS, Jain AK. A survey of automated visual inspection. Computer Vision and Image Understanding. 1995; **61**(2):231-262

[8] Yeow JA, Ng PK, Tan KS, Chin TS, Lim WY. Effects of stress, repetition, fatigue and work environment on human error in manufacturing industries. Journal of Applied Sciences. 2014;**14**(24):3464-3347

[9] See J. Visual Inspection: A Review of the Literature [Technical Report]. Albuquerque, NM, USA: Sandia National Laboratories; 2012

[10] Cui Y, Liu Z, Lian S. A survey on unsupervised visual industrial anomaly detection algorithms. arXiv. New Jersey, USA; 2022. pp. 55297-55315

[11] Liu J, Xie G, Wang J, Li S, Wang C, Zheng F, et al. Deep industrial image anomaly detection: A survey. Machine Intelligence Research (Springer Science and Business Media LLC). 2024;**21**(1): 104-135. DOI: 10.1007/s11633-023-1459-z. ISSN 2731-5398

[12] Tao X, Gong X, Zhang X, Yan S, Adak C. Deep learning for unsupervised anomaly localization in industrial images: A survey. IEEE Transactions on Instrumentation and Measurement. 2022;**71**:1-21

[13] Lehr J, Sargsyan A, Pape M, Philipps J, Krüger J. Automated optical inspection using anomaly detection and unsupervised defect clustering. In: 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). Vol. 1. New Jersey, NJ, USA: IEEE; 2020. pp. 1235-1238

[14] Rippel O, Merhof D. Anomaly detection for automated visual inspection: A review. Bildverarbeitung in der Automation: Ausgewählte Beiträge des Jahreskolloquiums BVAu. 2023;**2022**: 1-13

[15] Su B, Zhou Z, Chen H. Pvel-ad: A large-scale open-world dataset for photovoltaic cell anomaly detection. IEEE Transactions on Industrial Informatics. 2023;**19**(1):404-413

[16] Zavrtanik V, Kristan M, Skočaj D. DRÆM – A discriminatively trained reconstruction embedding for surface anomaly detection. arXiv. Clinical Orthopaedics and Related Research. 2021;**abs/2108.07610**. Available from: https://arxiv.org/abs/2108.07610

[17] Ristea N-C, Madan N, Ionescu RT, Nasrollahi K, Khan FS, Moeslund TB, et al. Self-supervised predictive convolutional attentive block for anomaly detection. arXiv. Clinical Orthopaedics and Related Research. 2021;**abs/2111.09099**. Available from: https://arxiv.org/abs/2111.09099

[18] Rippel O, Mertens P, Konig E, Merhof D. Gaussian anomaly detection by modeling the distribution of normal data in pretrained deep features. IEEE Transactions on Instrumentation and Measurement. 2021;**70**:1-13

[19] Cohen N, Hoshen Y. Sub-image anomaly detection with deep pyramid correspondences. arXiv. Clinical Orthopaedics and Related Research. 2020;**abs/2005.02357**. Available from: https://arxiv.org/abs/2005.02357

[20] Defard T, Setkov A, Loesch A, Audigier R. PaDiM: a patch distribution modeling framework for anomaly detection and localization. arXiv. 2020

[21] Roth K, Pemula L, Zepeda J, Schölkopf B, Brox T, Gehler PV. Towards total recall in industrial anomaly detection. arXiv. Clinical Orthopaedics and Related Research. 2021;**abs/2106.08265**. Available from: https://arxiv.org/abs/2106.08265

[22] Li N, Jiang K, Ma Z, Wei X, Hong X, Gong Y. Anomaly detection via self-organizing map. arXiv. Clinical Orthopaedics and Related Research. 2021;**abs/2107.09903**. Available from: https://arxiv.org/abs/2107.09903

[23] Ishida K, Takena Y, Nota Y, Mochizuki R, Matsumura I, Ohashi G. Sa-patchcore: Anomaly detection in dataset with co-occurrence relationships using self-attention. IEEE Access. 2023; **11**:3232-3240

[24] Kim J-H, Kim D-H, Yi S, Lee T. Semi-orthogonal embedding for efficient unsupervised anomaly segmentation. arXiv. Clinical Orthopaedics and Related Research. 2021;**abs/2105.14737**. Available from: https://arxiv.org/abs/2105.14737

[25] Wang G, Han S, Ding E, Huang D. Student-teacher feature pyramid matching for unsupervised anomaly detection. arXiv. Clinical Orthopaedics and Related Research. 2021;**abs/2103.04257**. Available from: https://arxiv.org/abs/2103.04257

[26] Wirth R, Hipp J. CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. Vol. 1. Manchester; 2000. pp. 29-39

[27] Schiffer S, Rothermel AM, Ferrein A, Rosenthal-von der Pütten A. Look: AI at work! - analysing key aspects of AI-support at the work place. In: Yamshchikov I, Meißner P, Rezagholi S, editors. Workshop on Human-Machine Interaction (HUMAIN) held at KI 2024. Würzburg, Germany: Technical University of Applied Sciences Würzburg-Schweinfurt; 2024

[28] Breunig MM, Kriegel HP, Ng RT, Sander J. Lof: Identifying density-based

local outliers. In: ACM SIGMOD Conference. New York, NY, USA: Association for Computing Machinery/ ACM; 2000

[29] Schölkopf B, Williamson RC, Smola A, Shawe-Taylor J, Platt J. Support vector method for novelty detection. In: Solla S, Leen T, Müller K, editors. Advances in Neural Information Processing Systems. Vol. 12. Cambridge, MA, USA: MIT Press; 1999

[30] Liu FT, Ting KM, Zhou Z-H. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data. 2012;**6**(3):1-39

[31] Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNS are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv. Clinical Orthopaedics and Related Research. 2018;**abs/1811.12231**. Available from: http://arxiv.org/abs/ 1811.12231

[32] Harlacher M, Altepost A, Elsen I, Ferrein A, Hansen-Ampah A, Merx W, et al. Approach for the identification of requirements on the design of AI-supported work systems (in problem-based projects). In: AI in Business and Economics. Berlin: De Gruyter; 2023. pp. 87-99

[33] Dempster A, Petitjean F, Webb GI. ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. arXiv. Clinical Orthopaedics and Related Research. 2019;**abs/1910.13051**. Available from: http://arxiv.org/abs/ 1910.13051

Chapter 5

Predicting Exoplanets Habitability: Metrics and Models

Yash Patel

Abstract

The search for habitable exoplanets has become a critical focus in the field of astrobiology and planetary science, with technological advancements such as the Kepler and TESS missions enabling the discovery of thousands of exoplanets. However, determining whether these planets are habitable remains a significant challenge due to the complex factors that govern planetary habitability, including atmospheric composition, temperature, and distance from the host star. To address these challenges, several habitability metrics, such as the Earth Habitability Index (EHI) and the Cobb-Douglas Habitability Score have been developed to quantify a planet's potential to support life. Traditional machine learning models, while effective in classifying planets, often struggle with highly imbalanced datasets and tend to produce misleading results. This chapter explores the potential of deep learning techniques, particularly a variational auto-encoder (VAE)-based anomaly detection model, in predicting the habitability of exoplanets. By analyzing a broader range of planetary and stellar features, this unsupervised approach identifies anomalies, or rare habitable planets, in vast datasets. The proposed model's demonstrate that deep learning methods can provide a more accurate and nuanced approach to exoplanet habitability classification, overcoming the limitations of traditional methods. This chapter highlights the potential applications of anomaly detection techniques in astrobiology and their significance in the ongoing quest to find life beyond Earth.

Keywords: anomaly detection, exoplanets, habitability, deep learning, semi-supervised learning, predictive models, star systems

1. Introduction

The observable universe has an estimated radius of 46.5 billion light years and is believed to contain between 200 billion and 2 trillion celestial systems [1, 2]. The Milky Way, our own home galaxy, is just one of these systems. According to estimates from cosmologists, the Milky Way alone contains around 250 ± 150 billion stars, with stars similar to our Sun making up approximately 10% of this total, or around 20 ± 15 billion stars. On March 7, 2009, NASA launched the Kepler Space Telescope as part of its Discovery Program to search for planets the size of Earth orbiting other star systems [3, 4].

The discovery of habitable exoplanets, or planets outside of our solar system, has been made possible with the advancements in technology, such as the radial velocity method (Redshift), transit method, and the micro-lensing method [5]. These methods have enabled the detection and characterization of exoplanets with great precision. However, the challenge of identifying habitable exoplanets remains a significant task.

The quest to find habitable exoplanets—planets orbiting stars outside our solar system—has captivated scientists and astronomers for decades. The idea that another Earth-like planet may exist in the vastness of space raises profound questions about the potential for extraterrestrial life and the future of humanity in the cosmos. Technological advances in space missions, such as NASA's Kepler and Transiting Exoplanet Survey Satellite (TESS), have resulted in the discovery of thousands of exoplanets across diverse star systems. However, the challenge lies in determining which of these distant worlds have the necessary conditions to support life.

At the heart of this search is the concept of *habitability*. This complex term refers to a planet's ability to sustain life as we know it, typically defined by factors such as the planet's distance from its star (which affects its temperature), its atmospheric composition, the presence of water, and more [6]. Earth serves as the standard against which all other planets are compared, giving rise to metrics like the Earth Habitability Index (EHI) and the Earth Similarity Index (ESI). These metrics aim to quantify a planet's potential to host life, but they come with their own limitations, especially when applied to vast datasets that are often imbalanced—where only a small percentage of known planets are categorized as potentially habitable.

The classification of exoplanets as habitable or non-habitable remains a significant challenge in the field of planetary science. Traditional machine learning models, including decision trees and support vector machines (SVM), have been used to tackle this problem, but they often fall short due to the highly imbalanced nature of the data. For instance, only a small fraction of known exoplanets exhibit conditions that might support life, which can lead to misleading accuracy scores in binary classification models. A model may achieve high overall accuracy but still fail to correctly identify the rare instances of habitable planets. This imbalance not only skews the results but also limits the effectiveness of these models in practical applications.

Recognizing these limitations, researchers have turned toward more sophisticated techniques like deep learning and anomaly detection. Deep learning models, particularly variational auto-encoders (VAE), offer a promising alternative by analyzing complex datasets and identifying patterns that traditional models might miss. Unlike binary classifiers, VAE's can detect subtle anomalies in the data—planets that differ significantly from the majority yet possess the characteristics necessary for habitability. By focusing on these outliers, we can improve our ability to identify potentially habitable exoplanets.

This chapter explores a novel approach to exoplanet habitability prediction using a VAE-based anomaly detection model. The key objectives include:

- *Proposing a deep learning-based model*: The model uses a variational auto-encoder to learn underlying patterns in exoplanet data, allowing it to detect anomalies that could indicate habitable planets.

- *Feature expansion*: The model considers a broader set of planetary and stellar features compared to previous models, providing a more comprehensive analysis of each planet's habitability potential.

- *Performance evaluation*: The chapter evaluates the model's performance using metrics such as the receiver operating characteristic (ROC) curve, precision, recall, and area under the curve (AUC) scores, and compares it with traditional machine learning approaches.

- *Future directions*: The chapter concludes by discussing potential applications of this methodology in future space missions and the implications of discovering habitable exoplanets.

The next section provides a detailed overview of the various metrics used to evaluate exoplanet habitability, including the Earth Habitability Index and the Cobb-Douglas Habitability Score. The subsequent section focuses on models used for predicting habitability, from traditional machine learning methods to advanced deep learning approaches. Following that, we introduce the proposed anomaly detection model based on the variational auto-encoder, along with its methodology and workflow. We conclude with a discussion of the results, including a performance comparison of different models, and explore the potential for future research in this domain.

2. Habitability metrics

Predicting the habitability of exoplanets involves analyzing several planetary and stellar features to determine whether these distant worlds possess conditions conducive to life. Researchers have developed various metrics to quantify a planet's potential to support life, considering factors such as atmospheric composition, temperature, distance from the host star, and more. These metrics are not only crucial for ranking planets by habitability but also for improving the efficiency of machine learning and deep learning models in classifying planets based on their likelihood to support life.

2.1 Earth habitability index (EHI)

The *Earth Habitability Index (EHI)* is one of the most widely used metrics to assess the habitability of exoplanets, with Earth serving as the reference point for comparison [1]. This index measures how similar an exoplanet is to Earth in terms of critical factors that are thought to contribute to life, such as surface temperature, atmospheric composition, and the presence of liquid water.

EHI compares these features of an exoplanet to the corresponding characteristics of Earth and provides a numerical score to indicate the planet's habitability potential. A higher EHI value suggests a planet is more Earth-like and, therefore, more likely to be habitable. EHI is particularly useful in initial surveys of exoplanets, as it allows astronomers to quickly filter through large datasets and identify promising candidates for further study.

However, the EHI has limitations. By using Earth as the only reference, it may overlook planets that could support life in environments different from Earth's. For example, life could potentially exist in extreme conditions—such as under high pressure or within thick ice—on planets that would score low on the EHI scale. Thus,

while EHI provides a useful starting point, it is insufficient on its own for a comprehensive assessment of exoplanet habitability.

2.2 Cobb-Douglas habitability score (CDHS)

The *Cobb-Douglas Habitability Score (CDHS)*, inspired by economic theory, is another important metric used to assess exoplanet habitability [1]. The CDHS is based on the *Cobb-Douglas production function*, which in economics is used to model the output of a system based on various inputs. In the context of planetary habitability, the CDHS uses a similar approach to combine various planetary and stellar features into a single habitability score.

The CDHS considers factors such as:

- *Stellar properties*: The mass, radius, and luminosity of the host star.
- *Planetary features*: The planet's radius, density, temperature, and distance from the host star.
- *Orbital characteristics*: The semi-major axis (distance between the planet and the star) and eccentricity (how elliptical the planet's orbit is).

One of the key advantages of the CDHS is that it provides a more flexible and comprehensive metric for habitability, as it accounts for a wider range of planetary and stellar characteristics compared to the EHI. Additionally, because it uses a production function framework, the CDHS allows for the possibility of different combinations of factors contributing to habitability. For example, a planet might be more massive than Earth but still habitable if other conditions—like distance from the star or atmospheric composition—are favorable.

Despite these strengths, the CDHS also has its challenges. It requires careful calibration of the function's parameters to ensure that the weights assigned to each feature reflect their actual importance in determining habitability. Furthermore, the function assumes that the relationship between these features is multiplicative, which may not always hold true in complex planetary systems.

2.3 Earth similarity index (ESI)

Another critical metric is the *Earth Similarity Index (ESI)*, which evaluates how similar an exoplanet is to Earth by considering multiple dimensions of planetary characteristics. While the ESI is often conflated with the EHI, they differ in methodology. The ESI focuses on specific parameters such as a planet's radius, density, escape velocity, and surface temperature. Each of these characteristics is compared to Earth's, and the degree of similarity is calculated using a weighted formula.

The ESI is particularly valuable in exoplanet habitability studies because it focuses on factors that are directly related to Earth-like life. Unlike the EHI, which gives a broad overview of habitability, the ESI provides a more focused analysis on planets that resemble Earth in size and temperature [7].

One of the most useful applications of the ESI is its ability to rank exoplanets based on their physical similarity to Earth. This makes it easier for scientists to prioritize planets for detailed study using space-based telescopes like the James Webb

Space Telescope. However, like the EHI, the ESI has its drawbacks—it is Earth-centric and thus might exclude planets that could support non-Earth-like life.

2.4 Planetary and stellar parameters

Several additional planetary and stellar parameters play a crucial role in determining habitability. While the specific metrics like EHI, CDHS, and ESI help summarize a planet's potential, they are all underpinned by detailed data about individual parameters. These parameters can be divided into two main categories:

- *Planetary parameters*:
    - *Mass and radius*: A planet's mass and radius influence its gravitational force, atmospheric retention, and geological activity. Larger planets may have stronger gravity, but they might also be gas giants unsuitable for life as we know it.
    - *Temperature and density*: These are key factors in determining whether a planet can support liquid water, one of the primary conditions for life. A planet's surface temperature depends on its distance from the star and its atmospheric composition.
    - *Atmosphere*: The presence of an atmosphere is essential for shielding the planet from harmful radiation and maintaining surface temperature. In particular, the concentration of greenhouse gases, such as CO_2, can influence surface conditions.
    - *Orbital eccentricity*: This determines the stability of a planet's orbit. Planets with highly elliptical orbits may experience extreme temperature variations, making it harder for stable life-supporting conditions to emerge.
- *Stellar parameters*:
    - *Star type and luminosity*: The type of star and its luminosity are critical for determining the habitable zone, the region around a star where liquid water can exist on a planet's surface. Cooler stars like red dwarfs have narrower habitable zones closer to the star, while hotter stars have wider zones that extend farther out.
    - *Stellar mass and age*: Younger stars tend to be more active and emit more harmful radiation, which could hinder the development of life. Older stars, on the other hand, may have more stable energy output, making them more conducive to life-supporting environments.
    - *Distance from the planet to the star (semi-major axis)*: This distance affects how much radiation a planet receives. Planets too close to the star may be too hot, while those too far may be too cold.

2.5 Challenges in applying habitability metrics

While the metrics discussed above offer valuable insights into exoplanet habitability, they are not without challenges. One of the primary difficulties lies in the *incompleteness of available data*. Many exoplanets are located thousands of light-years away, and the

observational data we have about them is often incomplete or uncertain. Key parameters like atmospheric composition, surface temperature, or even the planet's exact size may be missing, which complicates the task of calculating reliable habitability scores.

Another challenge is the *variability of star systems*. Stars come in a wide range of types, from cool red dwarfs to hot blue giants, and the conditions in their habitable zones can vary significantly. Planets orbiting red dwarfs, for instance, may experience extreme stellar flares, while those around more massive stars may have shorter lifespans due to the star's rapid evolution.

Finally, habitability metrics tend to focus on Earth-like conditions, which means they may overlook planets that could support life in very different environments. For example, moons of gas giants or planets with thick atmospheres could harbor life even if they do not fit neatly into the habitability frameworks currently in use.

The study of exoplanet habitability requires a multidimensional approach, and the development of metrics like the Earth Habitability Index, Cobb-Douglas Habitability Score, and Earth Similarity Index has been essential in advancing our understanding of which exoplanets might support life. These metrics rely on a variety of planetary and stellar parameters to quantify a planet's potential habitability, yet they are not without limitations. As future missions like PLATO and the James Webb Space Telescope provide more data, these metrics will evolve, and we may discover that life can thrive in environments far different from our own.

3. Models used for predicting habitability

The identification and classification of potentially habitable exoplanets require sophisticated techniques that can effectively analyze complex datasets. Over the years, various models have been developed, ranging from traditional machine learning algorithms to cutting-edge deep learning approaches. These models aim to classify exoplanets based on their potential to support life by analyzing planetary and stellar features. However, as the volume and complexity of exoplanet data have grown, limitations in traditional methods have become apparent, prompting the adoption of more advanced deep learning models. This section delves into the key models used in exoplanet habitability prediction, from early machine learning approaches to the latest anomaly detection techniques.

3.1 Traditional machine learning models

Machine learning has long been employed to address the problem of exoplanet classification and habitability prediction. Early approaches used well-established algorithms such as decision trees, support vector machines (SVMs), k-nearest neighbors (KNN), and logistic regression to process exoplanet datasets and classify planets as potentially habitable or non-habitable. These models rely on labeled data where exoplanets are categorized based on known metrics like the Earth Habitability Index (EHI) or Earth Similarity Index (ESI). While they provide useful insights, traditional machine learning models face several challenges, particularly in handling large, imbalanced datasets where only a small fraction of exoplanets are considered habitable.

3.1.1 Decision trees

Decision trees are simple yet powerful models used for classification tasks. In the context of exoplanet habitability, a decision tree works by recursively splitting the

data into subsets based on feature values, eventually assigning a label (habitable or non-habitable) to each exoplanet. This model is highly interpretable, as the decision-making process can be easily visualized in the form of a tree diagram [8].

However, decision trees tend to overfit the data, especially when dealing with noisy or imbalanced datasets, such as those common in exoplanet research. For example, only a small fraction of exoplanets in existing datasets are deemed potentially habitable, making it difficult for the decision tree to generalize effectively.

3.1.2 Support vector machines (SVM)

Support vector machines are another popular classification model used in exoplanet habitability prediction. SVMs work by finding a hyperplane that best separates the data points into different classes. In exoplanet classification, SVMs attempt to divide the dataset into habitable and non-habitable planets by maximizing the margin between the two classes [2, 3].

While SVMs are robust and effective for smaller datasets, they struggle with large, imbalanced datasets. The highly skewed nature of exoplanet datasets—where habitable planets make up a small minority—can lead to a high number of false negatives, where potentially habitable planets are misclassified as non-habitable. Additionally, SVMs require extensive feature engineering and parameter tuning, which can be a limitation when dealing with large, high-dimensional datasets.

3.1.3 K-nearest neighbors (KNN)

K-nearest neighbors is a simple, instance-based learning algorithm that classifies exoplanets based on their similarity to known habitable planets. Given a new exoplanet, the model searches for the K most similar planets in the dataset and assigns a label based on the majority class of these neighbors. While KNN is easy to implement and interpret, it suffers from significant performance degradation when applied to high-dimensional datasets like those used in exoplanet research [8–10].

One of the major limitations of KNN in habitability prediction is its sensitivity to the number of neighbors (K) and the choice of distance metric. Small changes in these parameters can drastically alter the model's performance. Moreover, KNN is computationally expensive, particularly when applied to large datasets, as it requires comparing each new data point to every other point in the dataset.

3.1.4 Logistic regression

Logistic regression, a linear model used for binary classification, has been employed to predict the habitability of exoplanets based on various planetary and stellar features. The model estimates the probability of a planet being habitable by fitting the data to a logistic curve. While it provides interpretable results and is computationally efficient, logistic regression assumes a linear relationship between the features and the target variable, which often does not hold in complex exoplanet datasets.

Additionally, logistic regression struggles with imbalanced datasets, leading to biased predictions toward the majority class (non-habitable planets). This limitation reduces its effectiveness in identifying potentially habitable planets in large, complex datasets.

3.1.5 Challenges of traditional machine learning

Despite their utility, traditional machine learning models face several challenges when applied to exoplanet habitability prediction:

- *Imbalanced datasets*: As mentioned earlier, most exoplanet datasets are highly imbalanced, with habitable planets forming a small percentage of the overall dataset. This imbalance often leads to poor performance in classification models, as the algorithms tend to favor the majority class.

- *Limited feature interactions*: Traditional models often struggle to capture complex, nonlinear interactions between planetary and stellar features, which are critical for accurate habitability prediction.

- *Overfitting*: Many machine learning models, especially decision trees, tend to overfit the training data, resulting in poor generalization to new, unseen exoplanet data.

Given these challenges, the field has increasingly shifted toward deep learning techniques that can handle larger datasets and more complex feature interactions.

3.2 Deep learning models

Deep learning models have emerged as a promising alternative to traditional machine learning approaches, particularly in tasks involving large, high-dimensional datasets like those used in exoplanet habitability prediction. Deep learning models, particularly neural networks, excel at automatically extracting meaningful patterns from raw data, making them well-suited for analyzing the complex relationships between planetary and stellar features.

3.2.1 Deep neural networks (DNN)

Deep neural networks consist of multiple layers of interconnected nodes, or neurons, that process input data to identify patterns and relationships. In the context of exoplanet habitability, DNNs can be trained on large datasets of exoplanetary features, learning to predict habitability by automatically identifying complex, nonlinear interactions between features like mass, radius, atmospheric composition, and stellar properties [11].

A notable example of deep neural networks in exoplanet research is the work by Rutuja Jagtap and colleagues, who used a deep convolutional neural network (CNN) based on the ASTRONET architecture to classify exoplanets as habitable or inhabitable [12]. By utilizing activation functions like ReLU and sigmoid, the network was able to capture intricate feature interactions and improve classification accuracy [8, 11].

3.2.2 Convolutional neural networks (CNN)

Convolutional neural networks, commonly used in image processing, have also been applied to exoplanet detection and classification. CNNs are particularly useful when analyzing time-series data or data with spatial relationships, such as planetary transits observed by space telescopes.

For instance, the work by Ishaani Priyadarshini and colleagues employed an ensemble CNN to classify exoplanets using data from NASA's Kepler mission. This approach involved training multiple CNNs on different subsets of the dataset and then combining their predictions to improve accuracy. CNNs are highly effective at capturing spatial and temporal patterns in the data, making them a powerful tool for detecting exoplanetary transits and predicting habitability [13].

3.2.3 Recurrent neural networks (RNN)

Recurrent neural networks (RNNs), which are designed to handle sequential data, have also been explored in exoplanet habitability prediction. RNNs are particularly useful for analyzing time-series data, such as changes in a planet's brightness over time or variations in its atmospheric composition.

By maintaining a memory of previous inputs, RNNs can capture temporal dependencies in the data, making them well-suited for predicting exoplanet habitability based on dynamic features like orbital eccentricity, surface temperature fluctuations, and stellar activity. However, RNNs are more challenging to train compared to other deep learning models, and their effectiveness in exoplanet habitability prediction is still an active area of research.

3.3 Anomaly detection techniques for habitability prediction

Anomaly detection techniques have gained attention in recent years as a means of overcoming the limitations of traditional classification models, particularly when dealing with imbalanced datasets. In exoplanet habitability prediction, anomaly detection models are designed to identify rare instances—that is, potentially habitable planets—by recognizing patterns that deviate from the majority of non-habitable planets.

3.3.1 Variational auto-encoders (VAE)

Variational auto-encoders (VAEs) are a powerful deep learning technique for anomaly detection, particularly in scenarios where labeled data is limited. A VAE is a type of unsupervised model that consists of two main components: an encoder and a decoder. The encoder compresses input data into a low-dimensional latent space, while the decoder reconstructs the original data from this latent space. By minimizing the reconstruction error, the VAE learns to identify unusual patterns or anomalies in the data, which can be interpreted as potentially habitable exoplanets [14].

In the context of exoplanet habitability prediction, a VAE-based model can analyze a wide range of planetary and stellar features, learning the typical characteristics of non-habitable planets and flagging planets with anomalous features as potentially habitable. Unlike traditional binary classifiers, which struggle with highly imbalanced datasets, VAEs excel at detecting rare, anomalous instances without requiring large amounts of labeled data.

3.3.2 Memetic algorithms

Memetic algorithms, a class of metaheuristic algorithms, have also been applied to exoplanet habitability prediction through anomaly detection. These algorithms combine global search techniques, such as genetic algorithms, with local optimization

methods to efficiently explore large, complex datasets. In habitability prediction, memetic algorithms have been used to cluster exoplanets based on their habitability features, identifying rare habitable planets as anomalies within the dataset [15].

A recent study by Jyotirmoy Sarkar et al. [15] introduced a novel multistage, multi-version memetic clustering algorithm for exoplanet habitability prediction. By applying anomaly detection to large exoplanetary datasets, the algorithm was able to identify potentially habitable planets that had been missed by traditional supervised learning models.

3.4 Comparative analysis of models

Both traditional and deep learning models have contributed significantly to the field of exoplanet habitability prediction, but each approach has its own strengths and limitations. Traditional machine learning models, while interpretable and easy to implement, often struggle with large, imbalanced datasets and fail to capture complex feature interactions. Deep learning models, particularly neural networks, excel at handling large datasets and automatically extracting meaningful patterns but require significant computational resources and training data.

Anomaly detection techniques, particularly VAEs and memetic algorithms, offer a promising alternative by focusing on the rare instances of habitable planets. These methods can handle imbalanced datasets more effectively and provide a more nuanced approach to habitability prediction by identifying anomalies that traditional classification models might miss.

The choice of model ultimately depends on the specific task at hand. For initial surveys of exoplanet habitability, traditional machine learning models may provide quick, interpretable results. However, for more detailed analysis and prediction, deep learning models and anomaly detection techniques offer greater accuracy and robustness, especially as more data from future space missions becomes available (**Table 1**).

The field of exoplanet habitability prediction has evolved significantly, with traditional machine learning models giving way to more advanced deep learning and anomaly detection techniques. While early models like decision trees, SVMs, and KNN provided valuable insights, their limitations in handling large, imbalanced datasets, and complex feature interactions prompted the adoption of deep learning methods like DNNs, CNNs, and RNNs. Furthermore, anomaly detection models, particularly VAEs and memetic algorithms, have shown great promise in identifying rare habitable planets in vast datasets. As new data continues to be collected from missions like PLATO and the James Webb Space Telescope, these advanced models will play an increasingly important role in the ongoing search for habitable exoplanets.

4. Proposed anomaly detection approach

The search for habitable exoplanets is particularly challenging due to the sheer volume of data and the significant imbalance between potentially habitable and non-habitable planets. Traditional machine learning models often struggle to classify habitable planets accurately because the vast majority of data points correspond to non-habitable exoplanets. To address these limitations, anomaly detection techniques, specifically *variational auto-encoders (VAE)*, provide a robust and efficient way to identify potentially habitable planets by treating them as rare anomalies in large

Model/Approach	Strengths	Weaknesses	Performance in exoplanet habitability
Decision trees	• Simple and interpretable • Fast training and prediction	• Prone to overfitting • Poor handling of imbalanced datasets	• Low recall for habitable planets • Struggles with complex data interactions
Random forests	• Reduces overfitting by averaging results • Handles large datasets better than decision trees	• Still struggles with imbalanced datasets • Requires hyperparameter tuning	• Moderate performance • Better than decision trees but low recall for rare habitable planets
Support vector machines (SVM)	• Effective in low-dimensional data • Can handle nonlinear relationships using kernels	• Sensitive to imbalanced data • Expensive to train on large datasets	• Difficulty in separating habitable from non-habitable planets in imbalanced data settings
Deep neural networks (DNN)	• Can model complex relationships • Scalable to large datasets	• Requires large labeled datasets • Prone to overfitting and low generalization on small/rare classes	• High accuracy on large datasets • Struggles to generalize to rare, habitable planets without extensive data augmentation
Convolutional neural networks (CNN)	• Excellent for spatial/time-series data • Learns spatial hierarchies effectively	• Not designed for anomaly detection • Computationally expensive • May overfit	• Works well with time-series data • Less suited for anomaly detection in habitability prediction
Auto-encoders (AEs)	• Good for dimensionality reduction and feature extractio • Can detect anomalies based on reconstruction error	• Lacks probabilistic interpretation in latent space • May struggle in noisy data	• Useful for anomaly detection but lacks robustness compared to VAE • Struggles with unseen or rare habitable planets
Variational auto-encoder (VAE)	• Effective for anomaly detection • Probabilistic latent space improves generalization • Handles imbalanced datasets • Unsupervised learning is scalable	• Sensitive to hyperparameters • More computationally complex than basic AEs	• High recall and precision for detecting habitable exoplanets as anomalies • Robust to unseen data due to probabilistic nature and regularized latent space

Model/Approach	Strengths	Weaknesses	Performance in exoplanet habitability
Memetic algorithm	• Combines global and local search techniques • Useful for handling large, complex datasets • Can improve optimization in feature space	• High computational cost • Sensitive to initial parameter settings	• Effective at clustering exoplanets based on habitability • Useful for optimizing multistage clustering but computationally expensive
Memetic algorithm with anomaly detection	• Incorporates anomaly detection in clustering • Balances exploration and exploitation in large datasets • Identifies rare habitable planets missed by traditional methods	• High computational demand • Tuning of multistage clustering parameters is complex	• Improved performance over traditional methods • Identifies habitable planets with higher accuracy due to multi-version memetic clustering

Table 1.
Comparison of the VAE-based anomaly detection approach with other state-of-the-art approaches.

datasets. This section outlines the detailed approach of using a VAE-based anomaly detection model for predicting exoplanet habitability [1].

4.1 Why anomaly detection?

Traditional classification models often perform poorly when faced with imbalanced datasets like those found in exoplanet research, where only a small fraction of the total dataset consists of habitable planets. In contrast, anomaly detection methods focus on identifying rare and unusual instances that differ significantly from the majority class, making them ideally suited for identifying potentially habitable exoplanets.

In the context of exoplanet habitability, anomaly detection aims to find planets that exhibit characteristics distinct from the majority of non-habitable planets. These anomalies may possess atmospheric, orbital, or stellar features that make them potential candidates for habitability. Anomaly detection offers two significant advantages:

- *Handling imbalanced data*: It is better suited to work with datasets where habitable planets are a minority.
- *Unsupervised learning*: It can detect anomalies without requiring extensive labeled data, which is useful when information about exoplanet habitability is incomplete (**Figure 1**).

4.2 Variational auto-encoder (VAE) overview

A *variational auto-encoder (VAE)* is an unsupervised deep learning model that is particularly effective for anomaly detection. It consists of two main components:

- *Encoder*: Compresses the input data into a lower-dimensional latent space.
- *Decoder*: Reconstructs the original data from the latent space representation.

The VAE works by learning a probability distribution in the latent space that captures the underlying structure of the input data. By minimizing the reconstruction error between the original data and the decoded output, the VAE learns to represent the typical patterns in the dataset. Exoplanets that deviate significantly from these learned patterns—due to unusual atmospheric composition, orbital characteristics, or other features—are flagged as anomalies, and these anomalies are considered potential candidates for habitability (**Figure 2**).

4.3 Dataset description

The dataset used in this approach is sourced from the *Planetary Habitability Laboratory's Exoplanet Catalog (PHL-EC)*, which contains detailed information on over 4500 exoplanets. This dataset includes both planetary and stellar features critical for determining habitability, such as [16, 17]:

- *Planetary features*: Mass, radius, density, surface temperature, atmospheric composition, escape velocity, orbital eccentricity, etc. (**Table 2**)

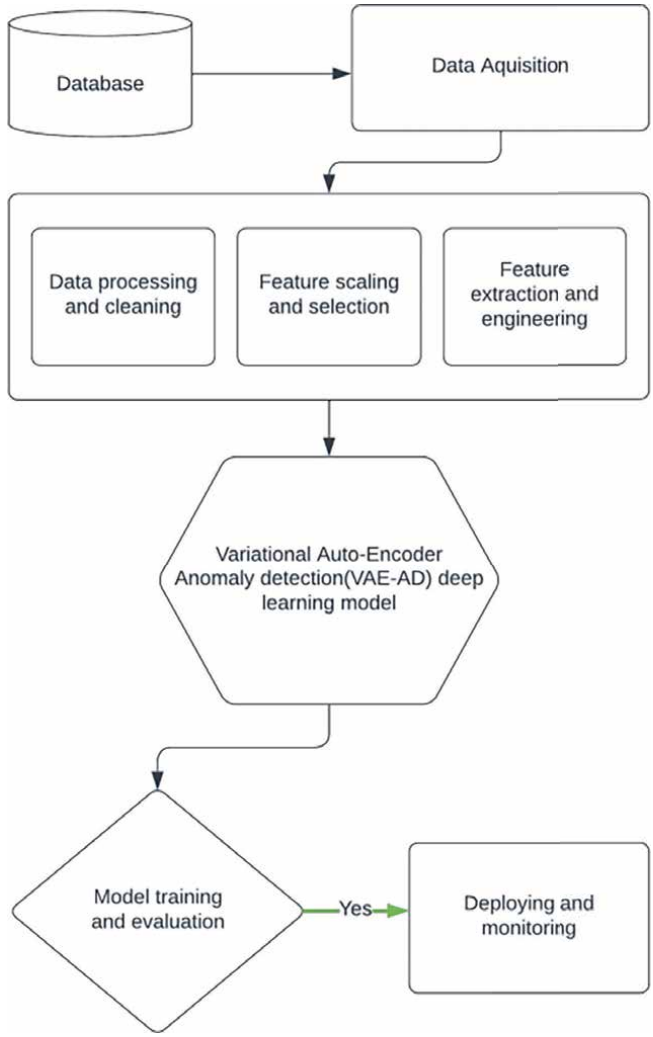

Figure 1.
Workflow of the proposed model [1].

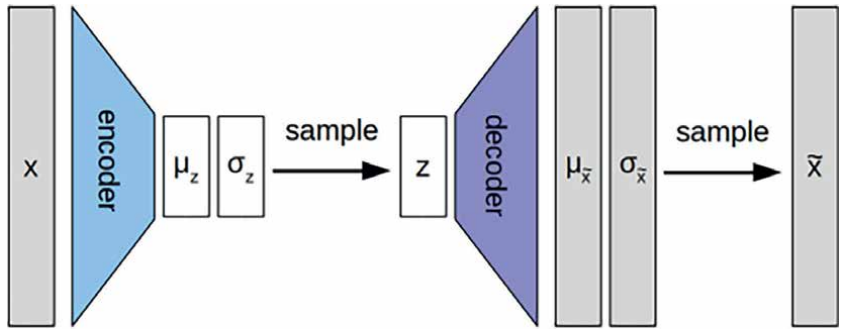

Figure 2.
Basic architecture for variational auto-encoders [15].

Planetary features
Composition Class, Atmosphere Class, MinMass(EU), Mass(EU), Radius (EU), Density (EU), Gravity (EU), EscVel (EU), SFluxMin, SFluxMean, SFluxMax, Teq Min (K), Teq Mean (K), Teq Max (K), SurfPress (EU), Mag, ApparSize (deg), Period (days), SemiMajorAxis (AU), Eccentricity, MeanDistance (AU), Omega (deg), ESI

Table 2.
List of all the planetary features used to train the model [1].

- *Stellar features*: Mass, radius, luminosity, effective temperature, metallicity, and distance from the host star (**Table 3**).

In total, the dataset includes 35 features (23 planetary features and 12 stellar features) that are relevant to the prediction of habitability. These features are preprocessed before being fed into the VAE model, which ensures that the data is normalized and free from missing values or outliers [18].

4.4 Data preprocessing

The preprocessing stage is essential to ensure that the VAE model operates effectively. Several steps are taken to clean and prepare the dataset:

- *Normalization*: All features are normalized to have a mean of zero and a unit variance. This ensures that features on different scales (e.g., mass, distance) are comparable.

- *Handling missing values*: Any missing values in the dataset are replaced using the mean value for that feature, ensuring that incomplete data does not affect the model's performance.

- *Outlier removal*: Outliers are removed or adjusted during preprocessing to prevent them from skewing the model's understanding of the data.

- *Feature scaling*: Scaling transforms the values of the features into a consistent range, making it easier for the VAE to learn the underlying patterns in the data.

The preprocessed dataset is then divided into two subsets: a *training set* (70%) and a *testing set* (30%). The training set is used to train the VAE model, while the test set is reserved for evaluating its performance on unseen data [19].

4.5 Model architecture

The *variational auto-encoder (VAE)* used in this approach consists of an encoder-decoder architecture designed to identify anomalies in the dataset. The model has two main components:

Stellar features
Mass (SU), Radius (SU), Teff (K), Luminosity (SU), S. [Fe/H], Age (Gyrs), ApparMag, Mag from Planet, Size from Planet (deg), HabZoneMin (AU), Hab Zone Max (AU)

Table 3.
List of all the stellar features used to train the model [1].

4.5.1 Encoder

The encoder compresses the input data into a lower-dimensional latent space. It consists of several fully connected layers that progressively reduce the dimensionality of the input. The encoder outputs two parameters, *mean* and *variance*, that define a Gaussian distribution in the latent space. These parameters allow the VAE to learn a smooth latent space, where each point represents a different exoplanet's features.

4.5.2 Latent space

The latent space is where the model represents the compressed version of the input data. This space allows the model to capture the underlying structure of the exoplanet dataset, helping it distinguish between typical exoplanets (non-habitable) and those with unusual, potentially habitable features.

4.5.3 Decoder

The decoder takes the latent space representation and reconstructs the original input data. It consists of fully connected layers that increase the dimensionality of the latent space until it matches the dimensions of the input. The goal of the decoder is to minimize the *reconstruction error*—the difference between the original data and the reconstructed data. This allows the model to learn which exoplanets fit the typical patterns in the dataset and which are anomalies.

4.5.4 Loss function

The VAE optimizes two loss components during training:

1. *Reconstruction loss*: Measures the difference between the input data and the reconstructed data. Exoplanets with large reconstruction errors are flagged as anomalies, as their features deviate significantly from the norm.

2. *Kullback-Leibler (KL) divergence loss*: Ensures that the latent space follows a standard normal distribution, making it easier for the model to generalize and detect anomalies.

For a VAE, the *total loss* is a combination of the *reconstruction loss* (which ensures that the decoder can accurately recreate the input from the latent space representation) and the *KL divergence loss* (which ensures that the latent space follows a normal distribution, typically $N(0,1)$.

4.5.5 KL divergence formula in VAE's

The KL divergence between two distributions $q(z|x)$ (the learned distribution of the latent variable z given the input x) and $p(z)$ (the prior distribution, usually a standard Gaussian $N(0,1)$) is given by:

$$D_{KL}\left(q(z|x)Pp(z)\right) = \int q(z|x)\log\left(q(z|x)/p(z)\right)dz \qquad (1)$$

For a VAE, assuming the encoder outputs a Gaussian distribution for the latent variable z, with mean μ and variance σ2, the KL divergence between this learned Gaussian N(μ,σ2) and the standard Gaussian N(0,1) can be simplified to the following closed-form expression:

$$D_{KL}\left(q(z|x)\|N(0,1)\right) = -(1/2)\sum_{1=2}^{d}\left(1+\log\left(\sigma_i^{\,2}\right)-\mu_i^{\,2}-\sigma_i^{\,2}\right) \quad (2)$$

Where:

- μ_i is the mean of the latent variable for dimension i,
- σ_i^2 is the variance (or $\log(\sigma_i^2)$ can be computed as the output of the encoder),
- d is the dimensionality of the latent space.

4.5.6 Total loss for VAE

The total loss for the VAE is:

$$\text{Total Loss} = \text{Reconstruction Loss} + \beta \cdot D_{KL}\left(q(z|x)\,P\,N(0,1)\right) \quad (3)$$

Where:

- *Reconstruction loss* measures the difference between the input x and the reconstructed output x^. It can be a mean squared error (MSE) or binary cross-entropy (BCE) depending on the type of data.
- *KL divergence loss* regularizes the latent space to match the prior distribution.
- β is a weighting factor in some VAE variants (such as β-VAE), which controls the trade-off between reconstruction and KL divergence loss.

4.6 Training and optimization

The VAE model is trained using the training dataset, where it learns to reconstruct the input data while minimizing the reconstruction and KL divergence losses. The model is trained for a set number of *epochs*, with each epoch representing one complete pass over the training dataset.

During training:

- *Batch size*: A small batch of data points is processed at a time to ensure efficient memory usage and faster training.
- *Optimizer*: The *Adam optimizer* is used to update the weights of the network, as it adapts the learning rate based on the gradient of the loss function.
- *Learning rate*: A low learning rate ensures that the model converges steadily, avoiding large updates that could lead to instability in training.

As the model trains, it learns to capture the typical patterns in the dataset, minimizing the reconstruction error for the majority of exoplanets (non-habitable planets). Any planets that generate a high reconstruction error are treated as anomalies, representing potential candidates for habitability.

4.7 Anomaly detection

Once the VAE model has been trained, it can be used to detect anomalies in the test dataset. The steps for anomaly detection are as follows:

1. *Reconstruction*: The VAE processes each exoplanet in the test set and attempts to reconstruct the original input features using the learned latent space representation.

2. *Reconstruction error calculation*: The reconstruction error for each exoplanet is calculated by comparing the original input to the reconstructed output. This error represents how well the model was able to replicate the exoplanet's features.

3. *Thresholding*: A threshold is applied to the reconstruction error. Exoplanets with a reconstruction error above the threshold are considered anomalies, as their features deviate significantly from the majority of non-habitable planets. These anomalies are flagged as potential candidates for habitability.

4.8 Performance evaluation

The performance of the VAE-based anomaly detection model is evaluated using several key metrics:

- *Receiver operating characteristic (ROC) curve*: The ROC curve plots the true positive rate against the false positive rate, providing a visual representation of the model's ability to distinguish between habitable and non-habitable exoplanets.

- *Area under the ROC curve (AUC)*: The AUC score quantifies the overall performance of the model, with a score closer to 1 indicating better performance.

- *Precision and recall*: Precision measures the proportion of true habitable planets among those flagged as anomalies, while recall measures the proportion of actual habitable planets detected by the model. These metrics help assess the model's effectiveness in identifying rare habitable exoplanets.

- *Confusion matrix*: The confusion matrix provides a detailed breakdown of the model's predictions, showing the number of true positives, false positives, true negatives, and false negatives.

4.9 Results

The proposed model variational auto-encoder (VAE) compared to other traditional machine learning and deep learning models overcomes the drawback of an imbalanced data set where binary classification is also highly imbalanced, with only 55 data points as true and the remaining 3734 data points categorized as false. Therefore, in order to overcome this problem of imbalanced data sets, we are exploring the field of anomaly detection in deep learning, more specifically variational auto-encoder (VAE's).

Figure 3 shows the plot between reconstruction error and the data points, which shows that the data points above a particular threshold error value can be regarded as an anomaly.

Figure 4 represents the confusion matrix of the predicted values from the testing data set. Here, the true label "0" represents the data point labeled as "Inhabitable," and

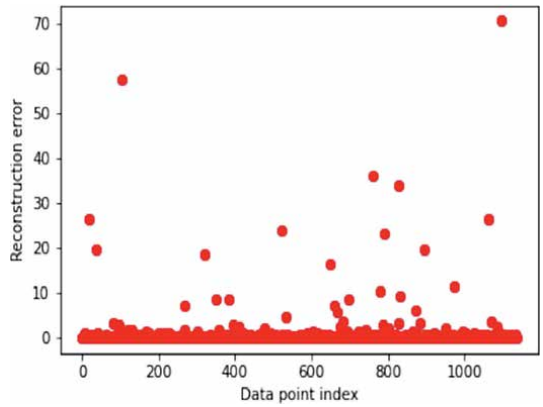

Figure 3.
Plot between reconstruction error and exoplanets [1].

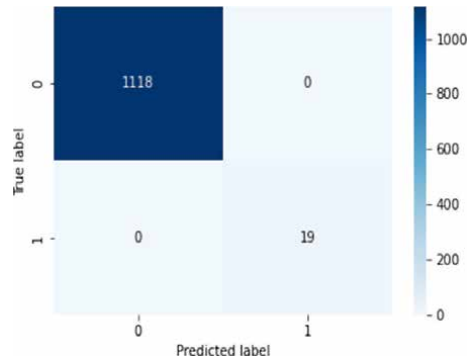

Figure 4.
Confusion matrix [1].

"1" represents the "Habitable" exoplanet. And as we can visualize from the matrix, 1118 exoplanets are correctly predicted as non-habitable and 19 exoplanets as habitable.

Figure 5 shows the receiver operating characteristic curve (ROC curve) and AUC score, which gives a better understanding of the model compared to the accuracy score because they take into account the balance of true positive rate (TPR) and false positive rate (FPR) in a classification model. In a binary classification problem, an accuracy score simply measures the percentage of correct predictions, regardless of whether they are true positives or true negatives. However, in some cases, it is more important to have a high TPR or a low FPR, such as in medical diagnoses where a false negative could be fatal.

In **Table 4**, we described the different AUC (area under the ROC curve) score for different threshold error and train-test split ratio for defining anomalies. From the AUC score table above, we can see that for the threshold value of 0.1 and the training and test split ratio of 70:30, the AUC score is 0.999246 which is relatively higher than other values for threshold error and train-test split ratio, which signifies it is a better model compared to the rest.

Figure 6 gives us a better understanding of the working of the proposed VAE model, which is a plot between mean square error (MSE) loss and epochs, which shows while training the model how training loss value is affected in relation to the number of epochs. As the number of epochs on training data increases, the MSE loss reduces.

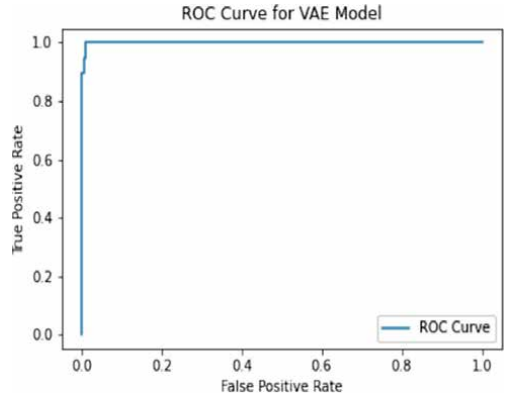

Figure 5.
ROC curve of testing data [1].

Threshold error value	Train test data split	AUC score
0.1	80:20	0.997627
0.05	80:20	0.997626
0.1	70:30	0.999246

Table 4.
AUC score for different threshold and train-test split [1].

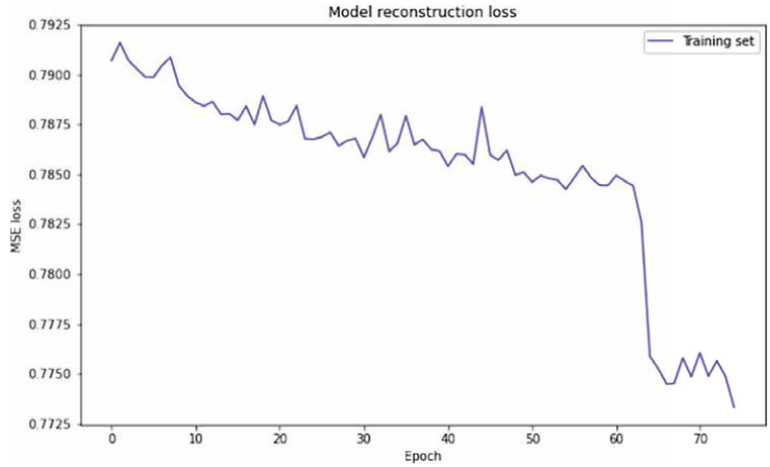

Figure 6.
Model reconstruction loss with increasing number of epochs [1].

The proposed VAE-based anomaly detection approach offers a powerful and efficient way to predict exoplanet habitability, overcoming the limitations of traditional machine learning models. By focusing on anomalies in the dataset, the VAE can identify rare, potentially habitable planets, even in the presence of highly imbalanced data. The model's unsupervised nature, combined with its ability to handle large, complex datasets, makes it an ideal tool for future space missions and the ongoing search for habitable exoplanets. As more data becomes available, this approach can be refined to improve its accuracy and potentially uncover new worlds capable of supporting life.

5. Future directions and conclusion

The exploration of exoplanet habitability is a rapidly evolving field, driven by advancements in both observational technology and data science methodologies. While significant strides have been made, there remain substantial challenges in accurately predicting which exoplanets might harbor life. The use of anomaly detection techniques, such as variational auto-encoders (VAE), has opened new avenues for tackling these challenges, particularly in dealing with the complexity and imbalance inherent in exoplanet datasets. As we look to the future, several promising directions could further enhance the accuracy and robustness of habitability predictions.

5.1 Integration of future space mission data

The ongoing and upcoming space missions, such as the *James Webb Space Telescope (JWST), PLAnetary Transits and Oscillations of stars (PLATO)*, and the *European Space Agency's (ESA) Ariel mission*, promise to deliver unprecedented volumes of high-quality data on exoplanets. These missions will provide more detailed observations of planetary atmospheres, surface conditions, and even potential bio-signatures, vastly improving the datasets available for habitability prediction. Integrating this new data into deep learning models will allow for more precise identification of habitable exoplanets.

Moreover, the application of *multi-wavelength observations* from these missions could improve the detection of specific atmospheric compounds (e.g., water vapor, methane, and carbon dioxide) that are key indicators of life. Combining data from different observational platforms will create richer datasets, offering deeper insights into planetary environments. Future anomaly detection models could be trained on these multidimensional datasets, enabling them to identify habitable planets with greater accuracy.

5.2 Improving model accuracy with data augmentation

One of the key challenges in exoplanet habitability prediction is the limited availability of labeled data for truly habitable planets. *Data augmentation techniques* could be used to address this issue by synthetically generating new data points that simulate different planetary conditions. This can help improve the training of deep learning models, particularly when real-world data is scarce. For example, by creating synthetic data based on variations in atmospheric composition, surface temperature, and orbital dynamics, the model could learn to generalize better, thereby improving its ability to detect habitable exoplanets from real observations.

Transfer learning is another approach that could enhance model performance. This technique involves training a model on a related task—such as planet classification based on atmospheric or geological features—and then fine-tuning it on the specific task of habitability prediction. By leveraging knowledge from other domains, the model can be more effective in identifying the nuanced features that indicate potential habitability.

5.3 Incorporating more complex planetary models

Current models primarily focus on key planetary and stellar features like mass, radius, temperature, and orbital characteristics. However, the discovery of *super-Earths, mini-Neptunes*, and other exotic planetary types suggests that habitability

may extend beyond the conditions found on Earth-like planets. Future models could benefit from incorporating more complex planetary physics, such as:

- *Planetary magnetic fields*: These play a crucial role in shielding a planet's atmosphere from stellar wind and radiation, which could be vital for maintaining habitability.
- *Geological activity and plate tectonics*: Active geology can help regulate atmospheric composition and surface temperature, both critical for sustaining life.
- *Exoplanet moons*: Large moons, such as those hypothesized to exist around gas giants, may themselves be potential sites for life. Models that account for tidal heating and orbital stability could better assess their habitability potential.

By broadening the range of planetary features considered, future anomaly detection models can more accurately predict the conditions necessary for habitability, even on non-Earth-like planets.

5.4 Exploring habitability beyond earth-like life

The current metrics and models for exoplanet habitability are largely Earth-centric, focusing on conditions that support life as we know it (e.g., liquid water, moderate temperatures). However, astrobiologists have long speculated about the possibility of *life in extreme environments*, such as organisms that can survive in acidic, high-pressure, or high-radiation environments. Future research could expand the scope of habitability models to explore these alternative forms of life, incorporating insights from extremophiles found in Earth's harshest environments (e.g., deep-sea hydrothermal vents or Antarctic subglacial lakes).

By developing models that account for a wider range of potential life-supporting conditions, scientists can expand the search for habitable exoplanets beyond Earth-like worlds, potentially discovering life in environments we currently consider uninhabitable.

5.5 Combining anomaly detection with explainable artificial intelligence (XAI)

As deep learning and anomaly detection models become more advanced, the complexity of these models often leads to a lack of transparency, making it difficult to interpret how predictions are made. This is particularly important when identifying habitable exoplanets, where scientific rigor and verifiability are critical. *Explainable AI (XAI)* techniques, which aim to make AI decision-making processes more transparent, can be integrated with anomaly detection to provide clearer insights into why a particular exoplanet was flagged as potentially habitable.

For example, XAI techniques like *Local Interpretable Model-Agnostic Explanations (LIME)* or *Shapley Additive Explanations (SHAP)* can highlight which features (e.g., atmospheric composition, orbital eccentricity) contributed most to the model's decision to classify an exoplanet as habitable. This would not only increase confidence in the model's predictions but also allow astronomers to focus their resources on planets that are most likely to support life.

5.6 Collaborative approaches and open data sharing

As the field of exoplanet habitability grows, collaboration across disciplines—including astronomy, data science, planetary geology, and astrobiology—will become

increasingly important. The development of *open-source platforms* for data sharing and model development can accelerate progress by allowing researchers worldwide to contribute to and refine existing models. Initiatives such as the *NASA Exoplanet Archive* and the *Planetary Habitability Laboratory (PHL)* already provide valuable datasets, but the future will likely see more collaborative platforms that allow for real-time data integration and model updates.

Crowd sourcing approaches, where citizen scientists or amateur astronomers contribute to exoplanet data analysis, could also play a role in expanding datasets and refining models. With more data and perspectives, anomaly detection models could continuously improve and evolve, leading to more accurate habitability predictions.

In this chapter, we explored the challenges and advancements in predicting exoplanet habitability using a variety of models and techniques, with a particular focus on *variational auto-encoder (VAE)-based anomaly detection*. Traditional machine learning methods, while effective to some extent, face limitations due to the highly imbalanced nature of exoplanet datasets and their inability to capture complex feature interactions. The VAE model offers a powerful alternative by focusing on the identification of rare, potentially habitable planets as anomalies in the dataset, overcoming these challenges.

The results of this approach are promising, providing a more nuanced way to identify habitable planets in large, complex datasets. By training the model on a wide range of planetary and stellar features, the VAE is able to detect subtle deviations that may indicate conditions favorable for life. Moreover, the model's ability to handle incomplete or imbalanced data makes it an ideal tool for analyzing the massive influx of exoplanetary data expected from future space missions.

However, the search for habitable exoplanets is far from complete. Future directions in this field include the integration of more comprehensive planetary models, the incorporation of multi-wavelength data from upcoming space missions, and the exploration of alternative forms of life. As we expand our understanding of what makes a planet habitable, we must also refine the models and metrics used to classify them, moving beyond Earth-centric definitions of habitability.

Ultimately, the development of advanced habitability prediction models will help guide the search for extraterrestrial life, allowing scientists to focus their resources on the most promising exoplanets. While we have yet to find definitive evidence of life beyond Earth, the continuous improvement of these models brings us closer to answering one of humanity's most profound questions: *Are we alone in the universe?*

Author details

Yash Patel
Indian Institute of Information Technology Allahabad, Prayagraj, India

*Address all correspondence to: yash.patel0313@gmail.com

References

[1] Patel Y, Tiwari S, Sonbhadra SK, Agarwal S. Predicting habitable exoplanets in different star-systems using deep learning based anomaly detection approach. In: Proceedings of 2023 International Joint Conference on Neural Networks (IJCNN); 18-22 June 2023; Gold Coast, Australia. IEEE; 2023. pp. 1-7. DOI: 10.1109/IJCNN54540.2023.10156122. Available from: https://ieeexplore.ieee.org/abstract/document/10191791

[2] Hora K. Classifying exoplanets as potentially habitable using machine learning. In: ICT Based Innovations: Proceedings of CSI 2015. Singapore: Springer; 2018. pp. 203-212. DOI: 10.1007/978-981-10-7566-7_18

[3] Mishra R. Predicting habitable exoplanets from NASA's Kepler mission data using machine learning [thesis]. 2017. (International Journal of Research in Engineering, Science and Management, 2020 (journal.ijresm.com))

[4] NASA Kepler Mission. Available from: https://www.nasa.gov/mission_pages/kepler/overview/index.html

[5] Sen S, Saha S, Chakraborty P, Singh KP. Implementation of neural network regression model for faster redshift analysis on cloud-based spark platform. In: Advances and Trends in Artificial Intelligence. From Theory to Practice: Proceedings of IEA/AIE 2021; 26-29 July 2021; Kuala Lumpur, Malaysia. Springer International Publishing; 2021. pp. 591-602. DOI: 10.1007/978-3-030-79457-6_50. Available from: https://link.springer.com/chapter/10.1007/978-3-030-79463-7_50

[6] Saha S, Nagaraj N, Mathur A, Yedida R, Sneha HR. Evolution of novel activation functions in neural network training for astronomy data: Habitability classification of exoplanets. The European Physical Journal Special Topics. 2020;**229**:2629-2738. DOI: 10.1140/epjst/e2020-000107-4

[7] Bora K, Saha S, Agrawal S, Safonova M, Routh S, Narasimhamurthy A. Cd-hpf: New habitability score via data analytic modeling. Astronomy and Computing. 2016;**17**:129-143. DOI: 10.1016/j.ascom.2016.09.003

[8] Jagtap R, Inamdar U, Dere S, Fatima M, Shardoor NB. Habitability of exoplanets using deep learning. In: Proceedings of IEEE IEMTRONICS 2021; 21-24 April 2021. IEEE; 2021. pp. 1-6. DOI: 10.1109/IEMTRONICS52119.2021.9422703. Available from: https://ieeexplore.ieee.org/abstract/document/9422571

[9] Rahman M, Afrin N. Finding habitable exo planets using boosting algorithm [thesis]. Dhaka: Brac University; 2018

[10] Singh SP, Misra DK. Exoplanet hunting in deep space with machine learning. International Journal of Research in Engineering, Science and Management (IJRESM). 2020;**3**(9):187-192

[11] Saha S, Basak S, Safonova M, Bora K, Agrawal S, Sarkar P, et al. Theoretical validation of potential habitability via analytical and boosted tree methods: An optimistic study on recently discovered exoplanets. Astronomy and Computing. 2018;**23**:141-150. DOI: 10.1016/j.ascom.2018.01.003

[12] Basak S, Saha S, Mathur A, Bora K, Makhija S, Safonova M, et al. CEESA

meets machine learning: A constant elasticity earth similarity approach to habitability and classification of exoplanets. Astronomy and Computing. 2020;**30**:100335. DOI: 10.1016/j.ascom.2019.100335

[13] Priyadarshini I, Puri V. A convolutional neural network (CNN) based ensemble model for exoplanet detection. Earth Science Informatics. 2021;**14**:735-747. DOI: 10.1007/s12145-021-00601-0

[14] Bechberger L. What Is a β-Variational Autoencoder? 2018. Available from: https://lucas-bechberger.de/2018/12/07/what-is-a-%CE%B2-variational-autoencoder/

[15] Sarkar J, Bhatia K, Saha S, Safonova M, Sarkar S. Postulating exoplanetary habitability via a novel anomaly detection method. Monthly Notices of the Royal Astronomical Society. 2022;**510**(4):6022-6032. DOI: 10.1093/mnras/stab3924

[16] NASA Exoplanet Archive. Available from: https://exoplanetarchive.ipac.caltech.edu/docs/data.html

[17] PHL-EC Catalog. Available from: http://phl.upr.edu/projects/habitable-exoplanets-catalog

[18] NASA Kepler Data Column Definition. Available from: https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html

[19] Sen S, Agarwal S, Chakraborty P, Singh KP. Astronomical big data processing using machine learning: A comprehensive review. Experimental Astronomy. 2022;**53**(1):1-43. DOI: 10.1007/s10686-021-09779-1

Edited by Miguel Delgado-Prieto

Anomalies are early whispers of malfunction, intrusion, or even discovery. *Anomaly Detection - Methods, Complexities, and Applications* offers a concise, practice-oriented guide to turning those whispers into actionable insights. Blending attention-enhanced supervision, ensemble learning, and variational auto-encoders with sector-specific experience from smart manufacturing, renewable-rich power grids, secure e-commerce platforms, and astrophysical observatories, the volume gives readers a panoramic view of data pre-processing, imbalance handling, explainability, edge deployment, and real-time analytics. By uniting algorithmic rigor with implementation detail, the book equips engineers, researchers, and graduate students to design resilient monitoring systems, reduce operational risk, and unlock new knowledge from large, noisy datasets.

Andries Engelbrecht, Artificial Intelligence Series Editor

ISSN 2633-1403
ISBN 978-1-83634-360-8
9 781836 343608

Published in London, UK

IntechOpen