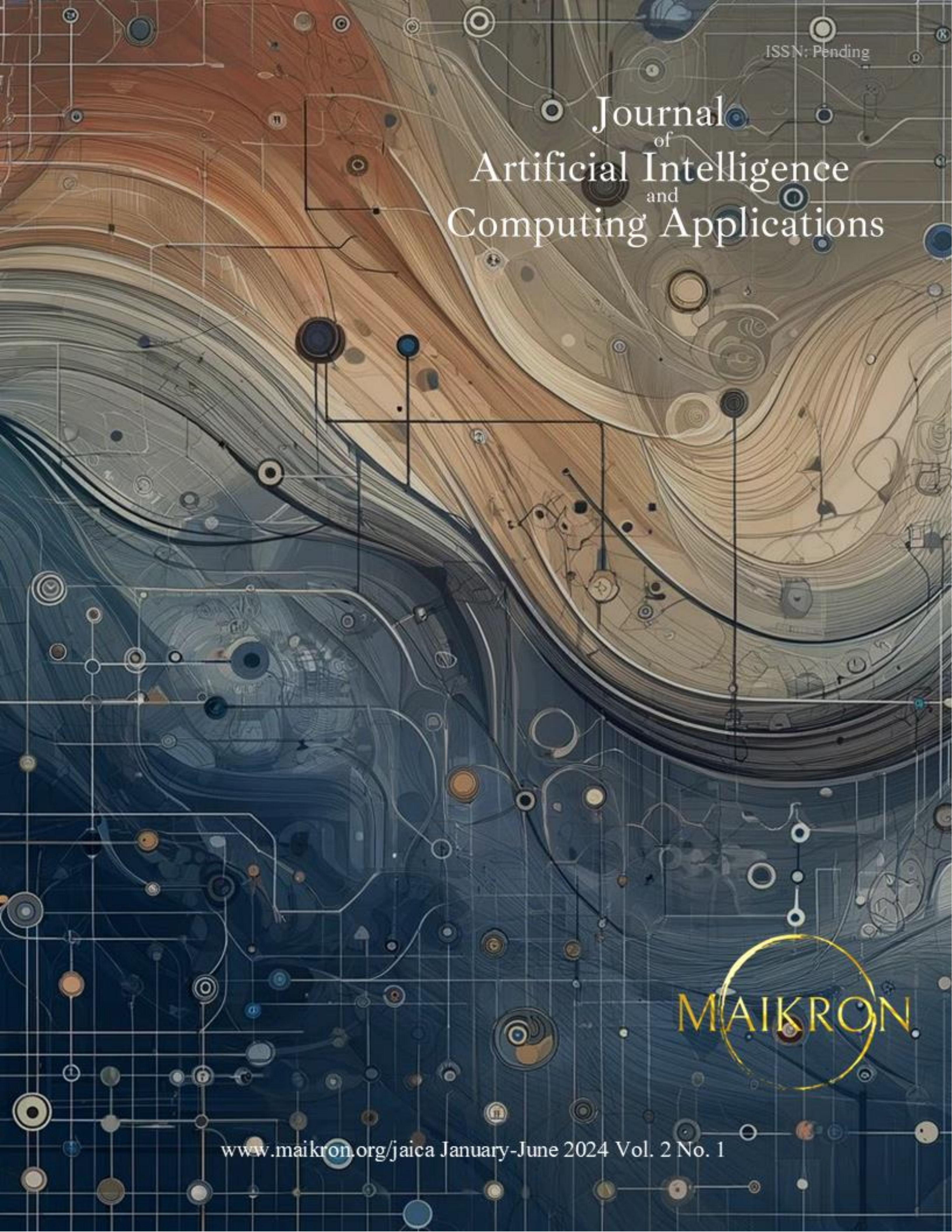
ISSN: Pending

Journal of Artificial Intelligence and Computing Applications



MAIKRON

www.maikron.org/jaica January-June 2024 Vol. 2 No. 1

Journal of **Artificial Intelligence** and **Computing Applications**

Journal of Artificial Intelligence and Computing Applications, Vol. 2, Núm. 1, enero – junio de 2024, es una publicación semestral editada por la Asociación para el Avance de las Aplicaciones Inteligentes y Tecnologías con Impacto Social, Calle 12 No. 103, Lotes No. 56 y 57, Ucú, Yucatán, C.P. 97357, Tel. (55) 28 55 70 20. URL: https://www.maikron.org/jaica, correo electrónico: contact@maikron.org. Editor responsable: Mauricio Gabriel Orozco del Castillo. Certificado de Reserva de Derechos al uso Exclusivo del Título: 04-2024-032513154000-102, eISSN: en trámite, por el Instituto Nacional del Derecho de Autor (INDAUTOR). Responsable de la última actualización de este número, Mauricio Gabriel Orozco del Castillo, Editor Responsable. Fecha de la última modificación: 30 de junio de 2024, Calle 12 No. 103, Lotes No. 56 y 57, C.P. 97357, Ucú, Yucatán. El contenido de los artículos es responsabilidad de los autores y no refleja el punto de vista de los árbitros, del Editor o de la asociación. Se autoriza la reproducción total o parcial de los textos siempre y cuando se cite la fuente completa y la dirección electrónica de la publicación.

Journal of Artificial Intelligence and Computing Applications
Vol. 2, No. 1, January-June 2024

Contents

Short Narrative Reviews

Critical Perspectives and Position Papers

Applied AI Exploration Papers

Journal of Artificial Intelligence and Computing Applications

MAIKRON

Foreword

Dear Readers, Contributors, and Colleagues,

It is with great pride that I introduce the second volume of the Journal of Artificial Intelligence and Computing Applications (JAICA). Building upon the momentum of our inaugural issue, this edition reaffirms our commitment to exploring the intersection of artificial intelligence and practical computing applications while expanding our vision with the introduction of new and thought-provoking contributions.

JAICA's mission extends beyond being a mere collection of scholarly articles. We aim to foster a dynamic community of researchers, practitioners, and enthusiasts united by a shared vision of advancing AI and computing technologies with meaningful, real-world impact. This vision is reflected in our inclusive approach, welcoming a wide array of contributions that range from comprehensive reviews to pioneering critiques and perspectives.

In this issue, we are pleased to present two insightful Short Narrative Reviews (SNRs). The first, titled "Machine Learning in Wireless Sensor Network Applications," delves into the integration of machine learning techniques in the domain of wireless sensor networks, highlighting both current trends and emerging challenges. The second article, "Convolutional Neural Networks for Identification of Forest Fires in Satellite Images," examines the application of convolutional neural networks (CNNs) in detecting forest fires, a critical area of research with significant environmental and societal implications.

Additionally, we are thrilled to debut our new section, Critical Perspectives and Position Papers (CPPP), which provides a platform for reflective, argument-driven discussions on AI's role in society. This section begins with two pioneering articles. The first, "Empathy, Accessibility, and Transparency in the Future of Artificial Intelligence: A Critical Perspective on Technology's Role in Modern Life," explores the dual nature of AI as a tool that can empower humanity while posing risks to creativity, community, and interpersonal connection. The second, "The eXplainable Artificial Intelligence Paradox in Law: Technological Limits and Legal

Transparency," critically examines the challenges of aligning AI-assisted legal systems with ethical and procedural standards, advocating for a balance between technical performance and societal trust.

We are also excited to present a contribution in our *Applied AI Exploration Papers (AAIEP)* section: *"Clustering-Based Cyber Situational Awareness: A Practical Approach for Masquerade Attack Detection."* This paper explores a clustering-based approach that integrates an optimized variant of K-Means with k-Nearest Neighbors to enhance intrusion detection in cybersecurity. By improving computational efficiency and integrating explainability mechanisms, this study provides valuable insights into AI-driven security frameworks and lays the groundwork for further research in cyber threat detection.

These articles collectively embody the spirit of JAICA by addressing pressing issues through rigorous analysis and creative exploration. The authors have skillfully navigated complex topics, offering valuable insights that contribute to our understanding of AI's role in solving real-world problems.

I extend my sincere appreciation to our dedicated editorial team, the diligent reviewers who provided invaluable feedback, and the supportive members of the AAAIMX student chapter. I am also grateful to our editorial sponsor, Maikron, for their unwavering commitment to advancing scholarly research in AI. Most importantly, I want to thank the authors for their trust in JAICA and for contributing their exceptional work to this issue.

Looking ahead, I am excited about the future of JAICA and the journey we are on. With each issue, we move closer to our vision of becoming a beacon of innovation and excellence in the fields of artificial intelligence and computing.

Warmest regards,

Mauricio G. Orozco-del-Castillo
Editor-in-Chief
Journal of Artificial Intelligence and Computing Applications (JAICA)

MAIKRON

Journal
of
Artificial Intelligence
and
Computing Applications

Review article

Machine learning in wireless sensor network applications: a short narrative review

Josué Pat-Cetina[1,*], Alejandro Pech-Escamilla[1], Teresita Chi-Pech[1], Halbert Eduardo Contreras-Villegas[1], and Daniel Visairo-Méndez[1]

[1]Tecnológico Nacional de México / IT de Mérida, Yucatán, México

ABSTRACT

This review explores the applications of Machine Learning in Wireless Sensor Networks, emphasizing its impact on various aspects such as routing, security, energy efficiency, speed, and quality. Its purpose is to bring attention to the most significant aspects and commonly employed applications of Machine Learning in Wireless Sensor networks for new and future research endeavors. The implications involved in obtaining 10 selected from 340 articles were the identification of specific articles, the screeening filtered by titles and abstracts and the Eligibility of the evaluated articles.The result of ten selected articles delve into the use of ML techniques, particularly Reinforcement Learning, with Q-learning being a prominent algorithm and so highlights the significance of ML in optimizing Wireless Sensor Networks performance, enhancing energy efficiency, and addressing specific challenges like wildfire detection and agricultural monitoring, systems that requires rapid response with low power consumption. Despite rigorous article selection, potential biases and criteria applicability limitations are acknowledged. Recommendations include further exploration of AI integration in practical applications, sophisticated approaches for energy optimization and security, and addressing emerging challenges in wireless sensor networks.

Keywords: wireless sensor networks, machine learning, deep learning

1. Introduction

Advancements in wireless technologies have been a major driver of modern human progress, enabling rapid, long-distance, and instantaneous communication. Among these technologies, the Internet of Things (IoT) stands out, integrating various devices and systems for seamless data exchange. At the core of many IoT systems are wireless sensor networks (WSNs), which play a crucial role in gathering and transmitting environmental data for real-time analysis.

WSNs are fundamental in applications such as battlefield surveillance, smart living environments, real-time environmental monitoring, and traffic optimization. By connecting numerous sensor nodes wirelessly, WSNs can measure various physical parameters precisely. However, efficient management of WSNs poses significant challenges, including issues related to routing, security, and, most critically, energy efficiency. Despite their design for low energy consumption [1], relying on batteries for WSNs with a large number of nodes is unsustainable due to the need for periodic replacement and the ecological risks associated with battery disposal [2].

Addressing energy efficiency is a priority, especially when utilizing renewable energy sources that may not provide a continuous power supply. Therefore, optimizing the energy consumption of WSNs is crucial, and this is where artificial intelligence (AI) and, more specif-

E-mail address: mg16081527@merida.tecnm.mx
https://doi.org/10.5281/zenodo.14165597

ically, machine learning (ML) methods, can be highly beneficial. ML techniques can enhance WSN performance by reducing energy usage and minimizing human intervention [3]. Additionally, in the context of data security, the increasing complexity of WSNs has highlighted vulnerabilities such as weak authentication, insecure network services, and poor encryption practices [4]. The lack of common standards balancing power consumption and security exacerbates these issues, leaving WSNs susceptible to attacks, especially in the communication layer [5].

The purpose of this review is to analyze the application of ML methods in WSNs, focusing on their role in enhancing energy efficiency and addressing security challenges. By examining the current state of research and anticipating future trends, this review aims to provide a comprehensive orientation of ongoing ML work in the context of WSNs.

This article is structured as follows: Section 2 outlines the inclusion and exclusion criteria used for selecting relevant studies, while Section 3 presents the main findings in terms of the different prevailing themes from the review of the selected articles. Section 4 discusses specific applications and benefits of ML in WSNs, highlighting deep learning (DL) techniques and their impact. Finally, Section 5 provides concluding remarks, summarizing the implications of ML in WSNs and projecting future challenges and research directions.

2. Methodology

The methodology of this review outlines a systematic approach aimed at identifying, selecting, and analyzing relevant literature on the applications of ML in WSNs. To ensure a comprehensive and high-quality selection of studies, a rigorous search strategy was implemented using defined inclusion and exclusion criteria. The process involved multiple stages, including initial identification of records, a thorough screening of titles and abstracts, an in-depth eligibility assessment based on quality metrics, and a final selection of the most impactful articles.

A total of 340 records were initially identified using Google Scholar with the search term intitle:"machine learning" AND intitle:"wireless sensor networks", covering publications from 2014 to the present. Following a preliminary review of the titles and abstracts, 43 articles were selected for an initial screening based on their relevance to the review topic, leaving 297 articles for a more detailed eligibility assessment.

During the eligibility phase, the full text of the remaining 297 articles was carefully examined. Of these, 67 articles were excluded as they addressed topics outside the scope of machine learning applications in wireless sensor networks. An additional 150 articles were removed because they were not indexed in the Journal Citation Reports (JCR), which was used as a quality filter to ensure the selection of impactful studies. Among the JCR-indexed articles, 55 were further excluded for not being classified within the top quartile (Q1), as the review prioritized high-impact research. Finally, 15 articles were excluded for failing to meet the threshold of an average citation rate of at least 1.5 citations per year.

The final selection consisted of 10 high-impact articles that provide a focused and comprehensive overview of ML applications in WSNs. This short narrative review includes a small number of studies, intentionally chosen to capture only the most impactful and influential works in the field. While this approach offers a concise yet in-depth analysis, it also introduces certain limitations. A broader, more exhaustive review could potentially include a wider range of studies, capturing additional insights and emerging trends that may have been omitted here. Nevertheless, the selection process ensured that the included articles represent significant contributions to the domain. This process, along with the stages of identification, screening, and eligibility assessment, is summarized in Figure 1. The complete list of the 10 selected articles is presented in Table 1.

3. Thematic Overview

The reviewed articles cover various applications of ML algorithms in WSNs, particularly those offering significant benefits tailored to specific objectives within the context of Industry 4.0. The use of ML techniques emerges as a strategic solution to key challenges in WSNs, including smart farming, routing, security, energy conservation, scheduling, localization, node clustering, data aggregation, fault detection, and real-time data integrity. DL techniques, integrated into automated networks, have shown promise, as highlighted in the work by [6]. These techniques are well-suited for handling dynamic, real-world scenarios through effective learning processes, reducing the need for frequent manual adjustments. This thematic overview summarizes and critically analyzes the contributions of ML algorithms to WSNs, identifying gaps and contradictions while maintaining a clear connection to the research objective stated in the Introduction.

In smart agriculture, ML techniques play a pivotal role in analyzing data collected by WSNs to identify various agricultural issues [8]. This application demonstrates the practical use of ML algorithms in addressing challenges within the agricultural domain. The integration of these technologies represents a progressive shift towards more advanced and sophisticated approaches for tackling agricultural concerns. By critically examining the literature, including the seminal work of [8], we highlight the current applications of ML in smart agriculture and identify areas for further exploration and improvement. This thematic analysis aligns with the research question presented in the Introduction, emphasizing the contemporary relevance and potential future developments of ML algorithms in the agricultural context.

The application of ML techniques in WSNs offers several advantages, including reduced computational complexity, improved feasibility in identifying optimal solutions, maximized resource utilization, extended net-

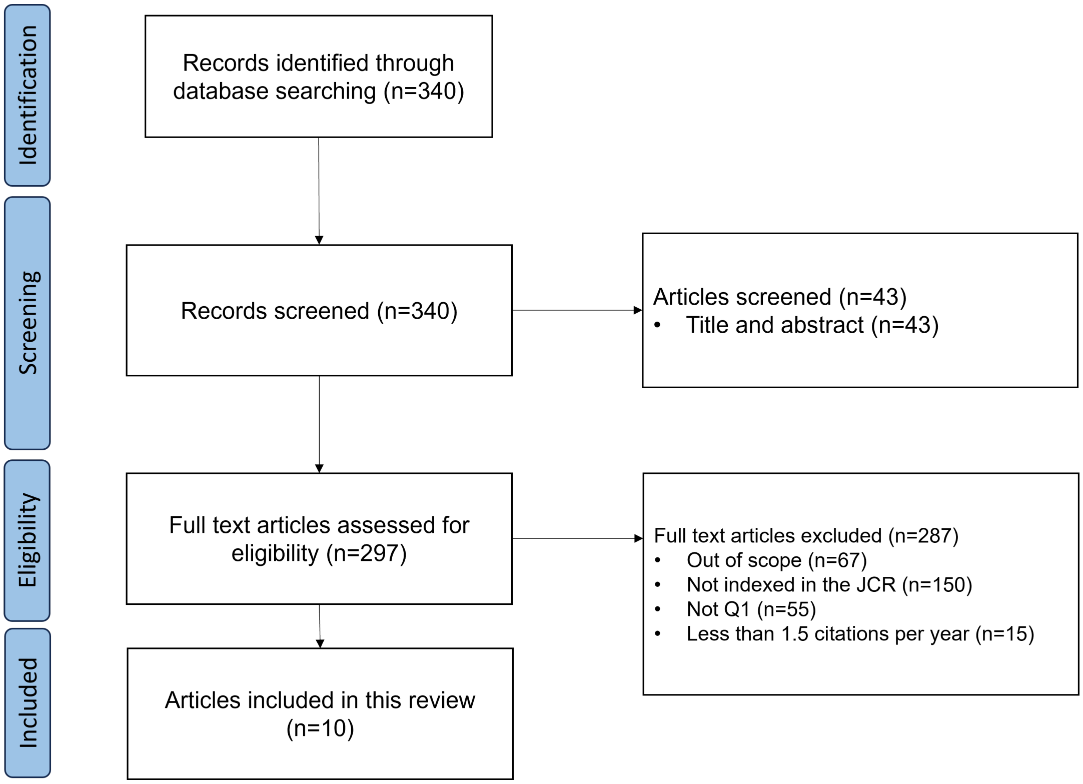

Identification
Screening
Eligibility
Included
Records identified through database searching (n=340)
Records screened (n=340)
Articles screened (n=43)
• Title and abstract (n=43)
Full text articles assessed for eligibility (n=297)
Full text articles excluded (n=287)
• Out of scope (n=67)
• Not indexed in the JCR (n=150)
• Not Q1 (n=55)
• Less than 1.5 citations per year (n=15)
Articles included in this review (n=10)

Figure 1. Diagram illustrating the systematic process of identifying, screening, and selecting relevant studies for this review on machine learning applications in wireless sensor networks. The diagram outlines each stage of the methodology, including the initial search, assessment of eligibility based on defined criteria, and the final inclusion of 10 high-impact articles.

work lifespan, and significant gains in energy efficiency [9]. A critical examination of the literature highlights the substantial contributions of ML to the field of WSNs, particularly in enhancing overall efficiency and extending the useful life of the networks.

Energy management is a crucial area where traditional methods, such as optimal routing protocols, node failure detection, and WSN topology construction, have been widely recognized for their utility. However, these methods have limited effectiveness in significantly reducing energy consumption. To overcome these limitations, advanced techniques like ML, metaheuristics, and Q-learning have been explored in the context of WSNs [7, 15].

The integration of IoT in WSNs has advanced significantly with the development of various smart devices. However, devices that are not connected to a conventional electrical grid face limitations in operating time. To address this, AI techniques, particularly reinforcement learning (RL), have been applied for energy optimization, providing enhanced control over battery charge levels and extending the operational lifespan of these devices [13]. The application of RL methods in WSNs also includes the identification of energy-efficient data transmission paths, which is essential for maintaining network efficiency. Additionally, RL techniques contribute to data security by mitigating risks associated with compromised nodes and potential data theft [12].

A practical application of DL is in combating forest fires, catastrophic events that cause significant economic, ecological, and environmental damage worldwide. Existing detection methods, such as satellite image processing systems, optical sensors, and digital cameras, often prove ineffective. In response, integrating WSNs with low-energy DL models has been proposed. By monitoring multiple sensor nodes measuring parameters like temperature, humidity, light intensity, and CO_2 levels, this approach stands out as a promising solution, enhancing detection efficiency and enabling rapid response to minimize potential damage [14].

Communication overloads in WSNs pose a significant challenge, as they can quickly deplete node energy and reduce network lifetime, thereby affecting the quality of service. To mitigate these issues, the Q-learning technique has been implemented. This RL approach involves two stages: (1) reward-dependent cluster head selection and (2) double-constraint path selection. By improving the efficiency of cluster head and route selection, Q-learning reduces uneven energy consumption and helps extend the network's lifetime [11].

Resilience in WSNs is another major challenge, re-

Table 1. The final list of articles used in this review, including information for title, journal, year of publication and citation.

Title	Journal	Year	Citation
Wireless sensor networks in agriculture through machine learning: A survey	Computers and Electronics in Agriculture	2022	[6]
A New Energy Prediction Algorithm for Energy-Harvesting Wireless Sensor Networks With Q-Learning	IEEE Access	2016	[7]
Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications	EEE Communications Surveys	2014	[8]
Machine Learning for Advanced Wireless Sensor Networks: A Review	IEEE Sensors Journal	2020	[9]
Resilient Routing Mechanism for Wireless Sensor Networks With Deep Learning Link Reliability Prediction	IEEE Access	2020	[10]
Optimizing the network energy of cloud assisted internet of things by using the adaptive neural learning approach in wireless sensor networks	Computers in Industry	2019	[11]
A Trusted Routing Scheme Using Blockchain and Reinforcement Learning for Wireless Sensor Networks	Sensors	2019	[12]
Reinforcement and deep reinforcement learning for wireless Internet of Things: A survey	Computer Communications	2019	[13]
Early Forest Fire Detection System using Wireless Sensor Network and Deep Learning	International Journal of Advanced Computer Science and Applications	2020	[14]
Machine Learning-Based Energy-Saving Framework for Environmental States-Adaptive Wireless Sensor Network	IEEE Access	2020	[15]

quiring the network to efficiently recover and adapt to changes in topology, such as node failures or attacks, especially when information about network links is incomplete. To address this issue, the Weighted Laplacian Deep Convolutional Neural Network (WL-DCNN) has been proposed. This DL model, combined with a predictive routing mechanism, optimizes data transmission and prolongs the network's useful life [10].

The thematic overview has highlighted several key areas where ML techniques have been effectively applied to enhance the performance of WSNs. From energy management and communication optimization to resilience and real-time applications, the integration of ML models, particularly DL and RL, shows great promise in addressing the inherent challenges of WSNs. This analysis provides a solid foundation for the subsequent discussion, where the implications and future directions of these advancements will be explored in greater detail.

4. Discussion

In WSNs, the speed and efficiency of data transmission are crucial considerations [14]. However, this efficiency must also account for security measures to protect the integrity of the transmitted data [5]. Balancing these factors becomes more complex when incorporating low-energy consumption nodes, which impose significant demands on both hardware and software, while still being constrained by energy limitations [12, 13]. This delicate balance between transmission efficiency, security, and power management poses a substantial challenge in the design and operation of WSNs, underscoring the need for innovative solutions that effectively address these competing requirements.

The variation in the findings of the reviewed studies was critically analyzed through a comprehensive comparison of the issues addressed in the selected articles. This analysis revealed common challenges consistently identified across multiple studies, including route optimization [13], energy-aware routing [8], energy saving [7, 15], security [10], speed [14], performance [12], quality of service [11], and programming complexity [9]. The goal of the review was to provide a detailed understanding of these variations and to identify patterns or consistencies in how the studies addressed key challenges within the scope of the selected articles.

To identify the most effective learning techniques for specific problems in WSNs, the findings of each applied method were interpreted within their respective contexts and evaluated based on their results. In WSN applications, these techniques have been utilized for continuous monitoring in agricultural fields [6], forest fire detection [14], and other scenarios. These techniques help improve routine processes and enable continuous monitoring, often replacing humans in tasks that may pose safety risks. However, to fully harness the potential of AI in WSNs, a large volume of data samples is required, which in turn demands significant time and resources [9].

RL has emerged as a key tool for energy optimization in WSNs, with Q-learning being identified as the preferred algorithm due to its simplicity and low computational complexity. Despite these advantages, the algorithm's impact on network performance has been significant, leading to notable improvements in transmission routes and the overall efficiency of WSN batteries, as demonstrated in studies such as [12, 13]. The Q-learning-based solar energy prediction (QL-SEP) algorithm is one notable approach that has shown promise in optimizing energy use in WSNs [7, 9, 11, 13, 15].

The Q-learning algorithm, a key component of RL, stands out as an essential tool in the field of ML. Its adaptive nature enables it to learn directly from interactions with the environment, making it applicable to a wide range of real-world problems. The algorithm's success has been demonstrated in practical applications, often outperforming human performance in various tests, highlighting its relevance and effectiveness in solving complex, dynamic problems. This RL approach is a valuable asset for addressing challenges that require optimized decision-making over time [9].

While RL techniques have shown promise in energy optimization, DL methods offer complementary advantages, particularly in event prediction and security applications. DL has proven effective for energy-aware routing in WSNs and has also been employed to address security issues [9, 10, 11, 12, 13]. Additionally, DL methods have been utilized in real-time event prediction tasks, providing valuable solutions for applications such as forest fire detection and continuous monitoring [14, 9, 11, 13].

However, it is essential to critically consider certain limitations in this review process. The thoroughness of the search strategy and the applicability of the inclusion and exclusion criteria may affect the representativeness of the selected studies. Additionally, attention should be given to the methodology used for assessing study quality and ensuring consistency in the application of criteria among reviewers. Although the review aims for transparency, the generalization of results and the temporal relevance of the information are important aspects to consider. Acknowledging these limitations is crucial for accurately interpreting the findings and identifying areas for future improvements in similar research endeavors.

The findings from the selected articles demonstrate a strong alignment with previous research, particularly in highlighting the importance of ML for optimizing the performance of WSNs and improving energy efficiency [9, 12, 13]. These results are consistent with earlier studies that emphasize the relevance of ML in the field of WSNs [12, 7]. Overall, the integration of ML techniques has proven effective in addressing core challenges such as energy management, data security, and resilience in WSNs. This synthesis of findings underscores the transformative potential of ML techniques in the ongoing development of WSNs, reinforcing their role as foundational components of smart, interconnected systems and paving the way for future research directions in this rapidly evolving field.

5. Conclusion

This review highlights the effective use of ML techniques in optimizing the performance of WSNs, with particular emphasis on RL and the widespread adoption of Q-learning as a prominent algorithm [9, 12, 13]. The findings demonstrate the significant impact of ML in enhancing energy efficiency and addressing critical challenges, such as wildfire detection [14] and agricultural monitoring [6], both of which require rapid response with low power consumption.

The review identified diverse applications of ML across key areas of WSNs, including routing, security, energy conservation, speed, and quality of service. It also explored the contributions of different ML algorithms, noting the importance of DL for energy-aware routing and improving WSN security. These insights underscore the transformative role of ML in advancing WSN capabilities and addressing complex, real-world challenges.

Despite the rigorous selection process, the review acknowledges certain limitations related to the applicability of inclusion and exclusion criteria and the potential for selection biases. Maintaining transparency in the methodology and ensuring consistency in evaluating study quality are crucial for accurately interpreting the findings and guiding future research efforts.

Further exploration is recommended in integrating AI into practical WSN applications, developing more sophisticated approaches for energy optimization and security, and tackling emerging challenges. The findings from this review serve as a valuable reference for future research, encouraging the pursuit of innovative and sustainable solutions in WSNs. By expanding the scope of literature searches to include not only Q1 but also high-quality Q2 articles, future studies can capture a broader range of emerging trends and insights, helping to guide new research initiatives and foster continued innovation.

This review demonstrates the substantial potential of ML techniques in driving the evolution and continuous improvement of WSNs within today's rapidly advancing technological landscape. The integration of ML into WSNs offers promising opportunities for enhanced efficiency, resilience, and intelligence, paving the way for more adaptive and capable networks in the future.

References

[1] Z. Sheng, C. Mahapatra, C. Zhu, and V. C. Leung, "Recent advances in industrial wireless sensor networks toward efficient management in iot," *IEEE Access*, vol. 3, pp. 622–637, 2015.

[2] S. Karnchanawong and P. Limpiteeprakan, “Evaluation of heavy metal leaching from spent household batteries disposed in municipal solid waste,” *Waste Management*, vol. 29, pp. 550–558, 2 2009.

[3] D. Praveen Kumar, T. Amgoth, and C. S. R. Annavarapu, “Machine learning algorithms for wireless sensor networks: A survey,” *Information Fusion*, vol. 49, pp. 1–25, sep 2019.

[4] A. Fotovvat, G. M. Rahman, S. S. Vedaei, and K. A. Wahid, “Comparative performance analysis of lightweight cryptography algorithms for iot sensor nodes,” *IEEE Internet of Things Journal*, vol. 8, 2021.

[5] S. Siboni, V. Sachidananda, Y. Meidan, M. Bohadana, Y. Mathov, S. Bhairav, A. Shabtai, and Y. Elovici, “Security testbed for internet-of-things devices,” *IEEE Transactions on Reliability*, vol. 68, 2019.

[6] M. M. Rahaman and M. Azharuddin, “Wireless sensor networks in agriculture through machine learning: A survey,” *Computers and Electronics in Agriculture*, vol. 197, p. 106928, 6 2022.

[7] S. Kosunalp, “A new energy prediction algorithm for energy-harvesting wireless sensor networks with q-learning,” *IEEE Access*, vol. 4, pp. 5755–5763, 2016.

[8] M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, “Machine learning in wireless sensor networks: Algorithms, strategies, and applications,” *IEEE Communications Surveys Tutorials*, vol. 16, pp. 1996–2018, 4 2014. [Online]. Available: https://ieeexplore.ieee.org/document/6805162

[9] T. Kim, L. F. Vecchietti, K. Choi, S. Lee, and D. Har, “Machine learning for advanced wireless sensor networks: A review,” *IEEE Sensors Journal*, vol. 21, pp. 12 379–12 397, 6 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9248049/

[10] R. Huang, L. Ma, G. Zhai, J. He, X. Chu, and H. Yan, “Resilient Routing Mechanism for Wireless Sensor Networks With Deep Learning Link Reliability Prediction,” *IEEE Access*, vol. 8, pp. 64 857–64 872, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9051677/

[11] A. Alarifi and A. Tolba, “Optimizing the network energy of cloud assisted internet of things by using the adaptive neural learning approach in wireless sensor networks,” *Computers in Industry*, vol. 106, pp. 133–141, apr 2019.

[12] J. Yang, S. He, Y. Xu, L. Chen, and J. Ren, “A trusted routing scheme using blockchain and reinforcement learning for wireless sensor networks,” *Sensors*, vol. 19, p. 970, 2 2019.

[13] M. S. Frikha, S. M. Gammar, A. Lahmadi, and L. Andrey, “Reinforcement and deep reinforcement learning for wireless internet of things: A survey,” *Computer Communications*, vol. 178, pp. 98–113, 10 2021.

[14] W. Benzekri, A. E. Moussati, O. Moussaoui, and M. Berrajaa, “Early forest fire detection system using wireless sensor network and deep learning,” *International Journal of Advanced Computer Science and Applications*, vol. 11, 2020.

[15] J. Kang, J. Kim, M. Kim, and M. Sohn, “Machine learning-based energy-saving framework for environmental states-adaptive wireless sensor network,” *IEEE Access*, vol. 8, 2020.

MAIKRON

Journal
of
Artificial Intelligence
and
Computing Applications

Review article

Convolutional neural networks for identification of forest fires in satellite images: a short narrative review

Randy Santos-Poot[1,*] **and Alejandro Alejandres-Rivera**[1]

[1]Tecnológico Nacional de México / IT de Mérida, Yucatán, México

ABSTRACT

Deforestation, a global phenomenon resulting in massive loss of forest areas, and forest fires, which are increasing in frequency and intensity due to climate change and human activity, present major challenges in managing and reducing these catastrophic events. Forests are essential for biodiversity and, representing about one third of the earth's land surface, require effective protection and conservation strategies as a matter of urgency. The effectiveness demonstrated by the models in detecting forest fires with satellite images is highlighted, allowing a faster response to emergencies. However, some limitations are pointed out, such as satellite capabilities and the need for high quality data to ensure the reliability of CNN model performance. This paper reviews recent advances in this field, highlighting the effectiveness of CNN-based models in identifying fires accurately and in a timely manner.

Keywords: forest fire detection, convolutional neural networks, multispectral satellite images

1. Introduction

Deforestation is a phenomenon that involves the massive loss of forest areas around the world, including forests, which are one of the most important natural resources for humanity. These represent almost a third of the planet's land surface and are home to countless species, especially in tropical areas, where there is usually greater biological diversity [1]. There are 24 areas made up of a large concentration of hotspots threatened by deforestation called "fronts", nine of which are in Latin America [2]. The increasing frequency and intensity of wildfires, exacerbated by climate change and human activity, pose significant challenges in the management and mitigation of these catastrophic events.

Convolutional Neural Networks (CNN) have proven to be promising tools in the field of computer vision and image analysis, particularly in the detection and classification of objects in images [3]. Their ability to learn relevant features automatically makes them ideal for forest fire detection tasks in images. The capacity of these technologies can not only save lives and minimize property damage, but also inform and guide long-term forest management strategies and conservation policies.

Despite significant advances in the application of CNN for wildfire detection in satellite images, significant gaps still remain in the literature. While research has demonstrated the effectiveness of lightweight CNN models in accurately identifying fires [3, 4], there is a lack of consensus on their performance under variable conditions, such as dense vegetation cover or the presence of clouds. Furthermore, the ability of these models to detect fires in incipient stages and their adaptability to different geographic environments also need to be further examined.

The objectives of this review article are multiple.

E-mail address: le20081650@merida.tecnm.mx
https://doi.org/10.5281/zenodo.14165778

Firstly, it seeks to synthesize and critically evaluate the existing literature on the performance and scope of CNN for the detection of forest fires in satellite images, highlighting both the significant advances and the identified limitations. The purpose is to identify specific areas of controversy and gaps in current research to highlight future research opportunities. In line with these objectives, the research questions focus on the effectiveness of CNN models under different environmental conditions, their ability to detect fires in early stages, and their adaptability to various geographic regions.

As this is a short narrative review article, a careful selection of high-quality articles was made, justified by the need to provide readers with a meaningful assessment of the current state of research on CNNs for fire detection. By focusing on the most influential works, a more direct and concise analysis of trends, challenges and opportunities in the field of wildfire detection using CNN can be provided.

In terms of structure, Section 2 will explain the process for selecting articles. Next, Section 3 will critically review the existing literature, highlighting the main findings, limitations and areas of controversy. Subsequently, Section 4 will discuss the implications of the review and possible future research directions in this emerging and vital field. Finally, in Section 5 the conclusions of the work are presented in a brief and concise manner.

2. Methodology

The methodology used in this review is essential to establish the scope, breadth and depth of the research. To ensure a comprehensive and rigorous search of the relevant literature, two search engines were mainly used: Google Scholar and Semantic Scholar. These platforms were selected due to their extensive repository of academic articles in various disciplines. Additionally, these platforms are recognized for their comprehensive indexing, user-friendly search capabilities, and integration with a wide range of academic journals.

The search began with the determination of the following keywords: “multispectral”, “satellite”, “images”, “convolutional”, “neural”, “networks”, “wildfire”, “detection” and “forest”. These keywords were used in combinations to form search terms using Boolean operators (only in the case of Google Scholar). Only those published within the last 5 years were considered, ensuring a balance between seminal works and contemporary ideas and the following search terms were used in the first week of March 2024:

1. “multispectral satellite images” AND “convolutional neural networks” AND “wildfire detection”, which returned 11 results in Google Scholar and 1530 in Semantic Scholar.

2. “multispectral satellite images” AND “convolutional neural networks” AND “forest fire detection” which returned 12 results in Google Scholar and 390 in Semantic Scholar.

In this way, a total result of 1922 articles was obtained in our search.

Inclusion criteria were strict to ensure a high standard of quality and relevance. The search was limited to journals indexed in the Journal Citation Reports (JCR), specifically those classified as Q1 or Q2, which represent higher quality in the field. Also, the scope was limited to works written in English. In addition, works were sought that reached at least 10 citations per year, reflecting their impact and recognition within the academic community. This approach ensured the inclusion of articles of high quality and relevance to the review.

The initial search returned a volume of 1922 articles between the two search terms, which were then filtered using the criteria mentioned above and shown in Figure 1. The approach was to filter these articles considering their relevance, citation count, and their quartile.

Of the 1922 initial articles, 1871 published articles whose title and abstract were not related to the topic were discarded, leaving 51. Then, 31 articles that did not belong to journals were discarded, leaving 20 articles. Of these, all were in English and were within the thematic scope of the research. One that had less than 10 citations per year and 5 that were not indexed in JCR were removed. This process reduced the list to 14 articles that met all inclusion criteria and were considered the most relevant and highest quality for the review. All these articles were published in journals belonging to Q1 or Q2. Table 1 shows the list of the final filtered articles.

The obvious limitation of the selection approach used was the rigor of the process, a decision made to maintain a simplified narrative and focus exclusively on the most innovative and highly cited research. While this methodology allows for a concise overview, it may omit certain relevant studies that did not meet strict criteria or that offer alternative or more nuanced perspectives. The potential for such omissions underscores the importance of the research, beyond what this review has encapsulated. Furthermore, an additional limitation was the large number of works in the results, which led to filtering only articles from the beginning, simplifying the process, but potentially omitting relevant studies that did not fit the search criteria. Despite these limitations, the applied methodologies and the selected literature constitute the foundation on which this work is based.

3. Thematic Overview

This analysis identifies several key themes within the detection of forest fires from images, with special attention to satellite data, due to its scope and large-scale monitoring capacity. Three main detection approaches are recognized based on the origin of the images used in the training models: ground, aerial and satellite detection. Although this article focuses on satellite detection, a brief review of ground and airborne approaches is also included to provide a comprehensive perspective.

Within satellite detection, studies are organized

Table 1. The final list of articles used in this review, including information for title, journal, year of publication and citation.

Title	Journal	Year	Citation
Wildfire detection using transfer learning on augmented datasets	Expert systems with applications	2020	[5]
A forest fire smoke detection model combining convolutional neural network and vision transformer	Frontiers in Forests and Global Change	2023	[6]
Forest fire detection in aerial vehicle videos using a deep ensemble neural network model	Aircraft Engineering and Aerospace Technology	2023	[7]
A Small Target Forest Fire Detection Model Based on YOLOv5 Improvement	Forests	2022	[8]
Forest-fire response system using deep-learning-based approaches with CCTV images and weather data	IEEE Access	2022	[9]
Active Fire Detection from Landsat-8 Imagery Using Deep Multiple Kernel Learning	Remote Sensing	2022	[10]
Active Fire Detection Using a Novel Convolutional Neural Network Based on Himawari-8 Satellite Images	Frontiers in Environmental Science	2022	[3]
Comparative Research on Forest Fire Image Segmentation Algorithms Based on Fully Convolutional Neural Networks	Forests	2022	[11]
A deep learning model using geostationary satellite data for forest fire detection with reduced detection latency	GIScience & Remote Sensing	2022	[12]
Super-Resolution Reconstruction of Remote Sensing Data Based on Multiple Satellite Sources for Forest Fire Smoke Segmentation	Remote Sensing	2023	[13]
Autonomous Satellite Wildfire Detection Using Hyperspectral Imagery and Neural Networks: A Case Study on Australian Wildfire	Remote Sensing	2023	[14]
Wildfire Detection Using Convolutional Neural Networks and PRISMA Hyperspectral Imagery A Spatial-Spectral Analysis	Remote Sensing	2024	[15]
Uni-temporal Sentinel-2 imagery for wildfire detection using deep learning semantic segmentation models	Geomatics, Natural Hazards and Risk	2023	[16]
Investigating the Impact of Using IR Bands on Early Fire Smoke Detection from Landsat Imagery with a Lightweight CNN Model	Remote Sensing	2022	[17]

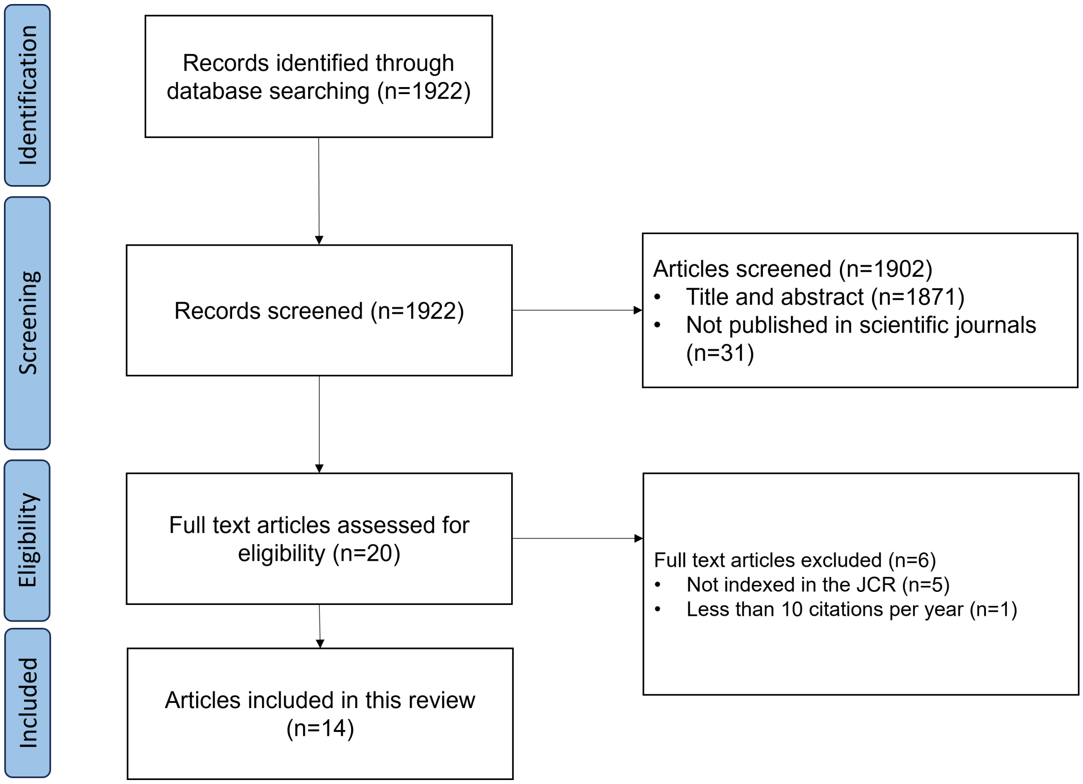

Identification
Screening
Eligibility
Included
Records identified through database searching (n=1922)
Records screened (n=1922)
Articles screened (n=1902)
• Title and abstract (n=1871)
• Not published in scientific journals (n=31)
Full text articles assessed for eligibility (n=20)
Full text articles excluded (n=6)
• Not indexed in the JCR (n=5)
• Less than 10 citations per year (n=1)
Articles included in this review (n=14)

Figure 1. Diagram of the selection process showing the different stages of identification, screening, eligibility and inclusion, filtering from 1922 articles to the final 14 considered in this review.

around three main themes: 1) development of active fire detection models using convolutional neural networks (CNN), which take advantage of satellite data of different spatial and temporal resolutions; 2) comparison and improvement of forest fire image segmentation algorithms, exploring the use of fully convolutional models for accurate fire and affected area detection; and 3) use of geostationary satellite data to minimize latency in fire detection, optimizing early response through high temporal resolution images. Additionally, advanced high-resolution remote sensing image reconstruction and segmentation methods for detecting smoke in wildfires, and the use of hyperspectral imagery to improve the accuracy of fire detection models, are discussed. These topics cover the most recent and sophisticated approaches to spatial and spectral analysis of data, contributing to the accuracy and reliability of detection.

3.1 Types of images for forest fire detection

Forest fire detection using images is divided into three forms depending on the source of the data: satellite, aerial and ground detection. Aerial images, taken from manned or unmanned aircraft (UAV), offer high spatial resolution, allowing specific areas to be observed in great detail. Its flexibility to fly over specific areas is an advantage, although its coverage is more limited compared to satellite images. Terrestrial imaging, on the other hand, allows for constant monitoring of the monitored area, as the cameras can operate without interruptions, although they are restricted to the places where they have been physically installed, thus providing a more limited range in relation to the others. types of images.

The satellite form of detection allows large areas of land to be covered, although, unlike terrestrial detection, it does not have fixed cameras to continuously monitor the same site due to the orbit of the satellites, which leaves an interval between images of the same place depending on the temporal resolution of the satellite. Satellites commonly used for creating datasets include Landsat-8, Himawari-8, PRISMA, VIIRS, and Sentinel-2. High spatial resolution satellites, such as Landsat-8 and Sentinel-2, allow smaller-scale fire detection, while high temporal resolution satellites, such as Himawari-8, facilitate early detection thanks to their orbits with reduced revisit time [10]. The satellites generate multispectral images (Landsat-8, Sentinel-2, VIIRS, Himawari-8) and also hyperspectral images (PRISMA). The difference between these types lies in the number of bands: multispectral images contain a small number of bands that cover wide ranges of the electromagnetic spectrum, while hyperspectral images, such as those from PRISMA, can contain up to 230 finer bands [15, 3, 10].

3.2 Moments of detection in the fire cycle

It is possible to define different approaches for fire detection, divided into three categories according to the time of the fire in which the detection is carried out: early, active and post-fire detection. Early detection is generally related to smoke detection, as satellite sensors often identify smoke before fire, since smoke spreads more quickly and fire detection is based on infrared bands. These bands allow the temperature of the Earth's surface to be captured, although the flames are required to have a minimum size, determined by the spatial resolution of the sensor [6]. Post-fire detection, on the other hand, focuses on identifying damaged areas after the fire.

3.3 Detection models based on Deep Learning

Deep learning (DL) is generally used as an improvement over threshold-based methods, as CNNs allow for more accurate pattern detection. Because wildfire detection using CNN is a recent approach, there are few large-scale datasets specifically designed for training fire detection models in the literature. Therefore, it is common for studies where models are proposed to generate their own datasets by labeling the phenomenon to be detected, such as smoke, fire, and burned areas.

These datasets are mostly designed for segmentation models, where detection is done at the pixel level, that is, each pixel of an image is labeled as a specific class, either in multiclasses or binary classes. There are also models where detection is done at the scene level [17, 6, 3, 10], so the label is applied to the entire image. Pixel-level detection allows you to observe the shape of the fire and its edges with greater precision. The input characteristics of the CNNs include reflectance values of the bands of each satellite, which vary depending on the radiometric resolution. Although the characteristics used in each model depend on the approach, in general spectral characteristics (band data) are used, and in some cases temporal characteristics (multitemporal band information) applicable to early detection. These datasets come from diverse regions, from small areas within a country [17, 16, 3] to global data sets with images of fires on multiple continents [10].

Regarding the architectures used, a wide variety of technologies and modules are observed, ranging from simple CNN architectures to hybrid models that combine CNN with Vision Transformer (ViT). Techniques such as multiscale convolution, residual edges, depth-separable convolution, inverted residual blocks, and spatial and channel attention modules have been used. Specialized architectures such as U-Net, ResNet, LinkNet, DeeplabV3, and Inception, among others, have also been identified. Models with lower parametric complexity, known as lightweight models, allow inferences to be carried out with lower latency, being suitable for early detection approaches [17, 6].

CNNs have limitations in their generalization capacity related with smoke [6], which has motivated the development of hybrid models such as SR-Net [6]. This model combines CNN and ViT to take advantage of both approaches and improve smoke detection. The training is carried out with true color images from the Himawari-8 satellite, which has a very high temporal resolution, ideal for early detection of fires. This model has been shown to be computationally more efficient and stable compared to other reference models, such as AlexNet, MobileNet, GoogLeNet and ResNet50, improving both adaptability and stability, evaluated using activation techniques such as Gradient-weighted Class Activation Mapping.

3.4 Optimization and improvement of detection models

The use of the visible spectrum bands (Red, Green, Blue) may not be sufficient to differentiate smoke from other aerosols in the atmosphere, which can lead to false alarms; Therefore, the use of infrared (IR) bands to improve the accuracy of smoke detection is common in the literature. It has been shown that the combined information from the IR bands can significantly increase the detection accuracy. An example is the lightweight VI_SD model, which performs scene-level classification [17], using a simplified but efficient structure, with spatial and channel attention modules that highlight relevant features. In addition, it incorporates residual modules to improve model learning. This model was compared with other state-of-the-art models, such as SmokeNet, SAFA, and Inception-ResNetV2, showing competitive performance with fewer parameters. When analyzing the contribution of the IR bands, it was found that the inclusion of the NIR band improved the model performance, while other band combinations yielded variations in the model accuracy.

[3] also uses images from the Himawari-8 satellite to demonstrate the potential of multiscale convolution and residual structures in fire detection efficiency. A CNN model called FireCNN is proposed, designed for real-time detection of forest fires. Classification is performed at the pixel level using a fully connected layer, similar to SimpleCNN, which processes and fuses features from multiple scale modules to calculate the probability that a pixel belongs to a fire, using Softmax. This approach optimizes the accuracy of wildfire detection by analyzing features at multiple scales, evidencing the effectiveness of FireCNN in accurate fire detection.

Multiscale convolution has been integrated into other architectures, such as Multiscale-Net, proposed in [10]. Multiscale-Net also uses different dilation rates and is trained with Landsat-8 images of various continents, allowing it to cover a variety of geographic scenarios and environmental conditions, ensuring the generalization and robustness of the model. Architectures like U-Net and FCN are also used in image segmentation; The encoder part of these architectures extracts features and generates lower resolution maps, while the decoder part produces segmentation masks. These masks integrate information from both intermediate and initial layers, combining local and global details to avoid the loss of important information and improve segmentation accuracy.

The inclusion of temporal information in model training has been shown to be relevant to improve early fire detection. The study in [12] explores latency reduction in wildfire detection using time series data from the Himawari-8 satellite, compared to MODIS and VIIRS. Three types of input features were evaluated: spectral, temporal and spatial, in regions of South Korea, North Korea and China. After classification, post-processing was applied to eliminate false alarms, using thresholds based on MODIS land cover ratios, helping to optimize the overall accuracy of the model.

Although Landsat-8 does not have high temporal resolution, its images are also useful in early detection, as the study by [13] shows. This work focuses on improving smoke detection in images from the VIIRS sensor, which has limited spatial resolution compared to sun-synchronous orbit satellites such as Landsat-8. To address this limitation, a CNN designed for super-resolution reconstruction of VIIRS images was used, in order to obtain high temporal and spatial resolution images. The CNN performs the super-resolution reconstruction, while the Smoke-Net network is responsible for segmenting the smoke in the enhanced images. Matched Landsat-8 and VIIRS images were used and groundtruth images were manually labeled. The VIIRS RGB images were adapted to the Landsat-8 RGB domain using a CycleGAN, allowing their reconstruction in super-resolution. The enhanced VIIRS images were then used for smoke segmentation with Smoke-Net, showing performance close to that of the original Landsat-8 images.

Detection latency plays a crucial role in early fire detection, and may depend on satellite orbit, model inference time, and image acquisition. To address this challenge, [14] proposes to obtain only sensing data from satellites instead of all sensor data. In this context, they developed and trained an optimized CNN for deployment on reliable autonomous satellites (TASO), adjusting its complexity to meet onboard processing requirements. A one-dimensional CNN (1-DCNN) proved to be effective for wildfire classification and, after evaluating various hardware options, the model was implemented on data processing devices on the satellite itself. This approach promises to significantly improve response capacity in natural disaster management.

3.5 Limitations and challenges in forest fire detection

Regarding the PRISMA satellite, there is a lack of suitable datasets for training segmentation models, which led [15] to manually create a dataset with the objective of evaluating the generalization capacity of four models. The models, which shared a basic structure of three hidden layers and an output layer with softmax for classification into seven classes, varied in complexity depending on their inputs. Although all models achieved good results in terms of accuracy, those that incorporated spatial information (neighboring pixels) in the input layer showed greater robustness and a lower false alarm rate.

Of the articles used in the review, the only recent study that has used data from the Sentinel-2 satellite is that of [16], in which several CNN models for semantic segmentation are compared. 14 models based on five encoder-decoder architectures were evaluated, including variations with encoders such as MobileNet and versions of ResNet. Among these, the U-Net model with ResNet50 as encoder showed outstanding performance and was selected for the creation of a pre-trained model using the Keras library. This model was evaluated with images of forest fires from several countries in climatic conditions similar to those of Turkey between 2021 and 2022, facing the challenge of adapting its performance to diverse conditions and the lack of specific datasets.

The accurate detection of forest fires is crucial for the prevention and mitigation of this natural disaster, the lack of datasets is a problem that not only affects satellite detection, this does not allow the training of sufficiently robust models. Although the images used for this type of detection are different from satellite images, which generally cover a large amount of space, they are designed to detect fires on a smaller scale and, therefore, are better in that regard than satellite detection but having limitations in the territory covered. In general, this field is relatively more advanced, such as the complexity of the architectures used, but they are more focused on active detection.

Two recent studies address this challenge from complementary perspectives. [5] highlights the scarcity of large-scale databases with real fire images, proposing an approach based on transfer learning combined with data augmentation techniques. This study used databases from Portugal and Corsica to retrain the Inception-v3 model, previously trained on ImageNet, in order to improve its performance in image-level fire detection using binary classification. Limitations were identified, including classification errors due to specific patterns and problems of reduced spatial scales, thus recommending a multiclass formulation and the consideration of consecutive frames to improve reliability. The second study [8], also uses transfer learning to overcome the limitation of a small dataset, focusing on improving the YOLOv5 model for real-time detection of forest fires. Tweaks include replacing the SPPF module with SPPFP to improve global information retention and adding a CBAM attention module, among other structural changes, optimizing multi-scale feature fusion. Although the system faces challenges such as false positives, proposed future optimizations could improve its real-time performance and applicability, especially for applications in drones or helicopter-mounted cameras.

A promising approach to improve fire detection accuracy is the use of deep neural network ensemble models, as proposed by [7]. In this study, an ensemble model was developed that combines four CNN models (Faster R-CNN, RetinaNet, Yolov2 and Yolov3) trained on the Corsica fire dataset, using data augmentation techniques to optimize performance. The model fuses the outputs of all four networks to improve fire detection accuracy, leveraging the strengths of each individual model. Although this approach notably improves accuracy, it was observed that each combination variant is more effective under certain specific conditions and that high computational complexity represents a major barrier to its real-time implementation on resource-limited platforms. However, the results suggest that the improvement in precision justifies the use of an ensemble model, especially in critical situations where early detection is key to preventing fire expansion.

Recently, the literature has begun to analyze in detail some cutting-edge architectures in fire detection, since these play a crucial role in extracting features from images, directly impacting the accuracy and efficiency of the model. Studies have shown that the choice of backbone can significantly influence model performance, as evidenced by research by [11] and [9]. In [11], a comparison was performed between four semantic segmentation networks to evaluate which is more effective in distinguishing between flame and forest background pixels in images captured by UAV. To improve generalization, data augmentation, random noise, and variations of environmental conditions were applied. The models evaluated included architectures such as FCN, U-Net, PSPNet and DeepLabV3+, tested with two backbones: VGG16 and ResNet50. Although similar performances were observed, the choice of ResNet50 as the backbone proved to be more effective overall, despite its higher computational complexity, which represents a challenge in terms of efficiency and applicability in resource-limited environments.

A modified Faster R-CNN architecture is proposed in [9], using DetNAS, a variant of Neural Architecture Search (NAS) for object detection networks. NAS is an automated method that searches the space of possible architectures to find the most suitable one for specific tasks, without relying exclusively on manual design. This study used DetNAS to identify an optimal backbone in the Faster R-CNN architecture, obtaining a lightweight model suitable for real-time detection. The resulting architecture was compared with others, such as ResNet, VoVNet and FBNetV3, the latter also based on NAS. The application of NAS for architectural optimization represents progress, although the model continues to face real-time implementation challenges in environments with limited computational resources.

As a conclusion to this section, the topics presented show a comprehensive overview of current techniques and approaches in forest fire detection using images. As part of the review of the different topics, the different types of images used (terrestrial, aerial and satellite), temporal detection approaches, as well as advances and improvements in the accuracy of the models through optimized architectures and the use of multispectral and hyperspectral bands. Although these techniques have allowed notable progress, challenges persist, particularly regarding the availability and quality of datasets, computational limitations, and the generalization capacity of models in different contexts and environmental conditions. These discussion points will be discussed further in the next section.

4. Discussion

In this study, a comprehensive literature review has been carried out to evaluate the performance and scope of CNNs in detecting forest fires using satellite images. In addition, we sought to identify the most effective model among those reviewed. The importance of this review lies in the growing interest in improving the accuracy and speed of fire detection, considering that prevention and early response are essential to mitigate the environmental and economic damage caused by forest fires. Although satellites offer significant advantages in coverage and cost, their spatial and temporal resolution can limit early detection, especially for smaller-scale fires. Satellite images, in particular, face the restriction of discontinuous coverage, where detection depends on the temporal resolution of the satellite and the sensitivity of its spectral bands. For example, detecting low-intensity fires may require satellites with high radiometric resolution, something that is not always available in current satellite orbits. This study also examines how DL techniques, such as CNNs, are beginning to overcome some of these limitations, proposing improvements in accuracy and responsiveness through advanced techniques such as spectral band combining and the use of hybrid architectures to address specific challenges of smaller-scale fires and in areas of high geographic diversity.

The results of the review indicate that CNNs have shown strong performance in fire detection at both the pixel and scene levels, as shown in Table 2. Although the models achieve good accuracy scores, these largely depend on the difficulty of the classification task and the characteristics of the dataset. The review reveals a clear limitation in terms of the lack of large-scale datasets, which prevents a robust evaluation of the generalization ability of the models. For example, in [15]'s study, the small size of the dataset, limited to images from high-biodiversity places like Australia, Sicily, and Oregon, resulted in a 20% drop between the test set metrics and the training set. This finding highlights the need for larger datasets that include greater geographic and seasonal variability for adequate evaluation and robustness of the models. Likewise, although no reduction in performance was observed in [14], the test set was composed of data obtained from a single image, which represents a considerable limitation, since it does not reflect all the variability of the patterns of forest fires in different environments and climates. These examples highlight the importance of expanding and diversifying datasets to improve the reliability of models in practice.

Among the high temporal resolution satellite datasets, those developed from Himawari-8 images stand out, with studies such as [6, 3, 12, 14] and [15] taking advantage of their high temporal frequency for the early detection. These sets offer the advantage of

Table 2. Comparison of technologies used in CNN models for forest fire detection.

Model	Metrics	Technology	Focus	Level
MultiScale-Net [10]	F1: 91%, IoU: 84.54%	U-Net, multiscale	Active, Fire	Pixel
SR-NET [6]	Recall: 97%, F1: 95%	CNN, ViT, multi-head attention	Early, Smoke	Scene
FireCNN [3]	Precision: 0.998, Recall: 0.999	Multiscale, Fully Connected	Active, Fire	Pixel
SN [12]	F1: 0.74	Classic CNN, Fully Connected	Early, Fire	Pixel
Smoke-Net [13]	IoU: 0.742	Residuals, attention, layer skip	Active, Smoke	Pixel
1-DCNN [14]	Accuracy: 97.83%	Classic CNN, Fully Connected	Damage/Active, Fire/Smoke	Pixel
3DCNN [15]	Average Accuracy: 0.68	3D Input, Classic CNN	Damage/Active, Fire/Smoke	Pixel
SN [16]	IoU: 97.98%	U-Net, ResNet50	Damage, Burned Area	Pixel
VIB_SD [17]	Accuracy: 93.57%	Spatial attention, channel, residuals	Early, Smoke	Scene

continuous coverage over time, although they present limitations in their spatial resolution that restrict their applicability in the detection of smaller-scale fires. In [6], 4,000 images of fires in China and Australia were used to classify between clouds and smoke, while the dataset in [3] focused on specific provinces in China, using the full disk product Himawari-8 L1. Preprocessing in [12], on the other hand, incorporated multiple thermal infrared bands (bands 7, 12, 13, 14, and 15) to improve accuracy by calculating the Brightness Temperature (BT). These variations in datasets reveal the need to choose resolution and features based on the specific geographic and temporal conditions of each fire scenario.

Higher spatial resolution datasets come from satellites such as Landsat-8 and Sentinel-2, and are especially useful for active detection and post-fire analysis. However, some studies have also used them in smoke-based early detection, such as in [13]. The latter is one of the largest sets, both in number of images and geographical coverage, although its temporal resolution limits its usefulness in real-time detection. In [17], an Australian-specific dataset was developed, based on Landsat multispectral imagery and processed by NBART, ideal for segmentation of burned areas. On the other hand, [16] focused on images of Türkiye with Sentinel-2, focused on evaluating post-fire effects through 21,690 images from 13 bands. These sets allow for detailed segmentation, but being geographically limited to regions such as Australia and Turkey, their global applicability is lower compared to global datasets such as [10]. The latter used Landsat-8 to build a global dataset, although its labels were generated algorithmically and not manually, which could affect the accuracy of the segmentation.

Both sets, being limited to specific regions such as Australia and Turkey, offer less global applicability compared to other datasets such as the one presented in [10], which takes advantage of the global coverage of Landsat-8 to develop a set of data spanning all continents. This set employs the SWIR2, SWIR1 and Blue bands, highlighting the sensitivity of SWIR2 to radiation emitted by fire, and employs the Active Fire Index to highlight fire areas while reducing smoke interference. However, this dataset was labeled using algorithms, in contrast to the manual labeling used in the other studies, which could affect the accuracy of pixel-level segmentation.

One of the common challenges in pixel-level datasets is class imbalance due to the size of fire-affected areas compared to the background. This makes it necessary to use specific metrics such as F1 score and IoU (Intersection over Union) to properly evaluate the performance of models on imbalanced datasets and precise segmentation.

In terms of architecture, most of the models reviewed use traditional CNN structures, and those that do not use them still achieve similar performance metrics, as seen in Table 3. This suggests that current architectures may be sufficient for the available datasets, although not necessarily optimized for complex conditions. Among the models analyzed, FireCNN and Multiscale-Net stand out. FireCNN, designed for real-time detection, achieved near 100% accuracy. However, this performance was evaluated in a very limited study area, centered on southern China, and was not validated in multiple contexts, raising questions about its generalizability. In contrast, Multiscale-Net, which used a large and diverse dataset (including images from different continents), achieved an IoU metric of 84.54%, demonstrating a high capacity to adapt to variability in fire patterns and environmental conditions. This architecture makes use of multi-scale dilation and convolution rates, allowing it to detect fires of different scales and improve its accuracy compared to standard CNN networks.

However, the effectiveness of these models remains limited by the quality and diversity of the available

Table 3. Datasets used in CNN models for forest fire detection.

Model	Period	Private	Region	Satellite	Size	Bands	Label
SR-Net [6]	2015-2022	Yes	CN, AU	Himawari-8	4,000	RGB	Manual, Smoke
Multiscale-Net [10]	2020	No	Global	Landsat-8	150,000	SWIR1, SWIR2, Blue	Algorithms, Active Fire
FireCNN [3]	2020	Yes	CN	Himawari-8	5,469	1-16	Manual, Active Fire
SN [12]	2015-2019	Yes	KR, CN	Himawari-8	2,157	7, 12, 13, 14, 15	Himawari Wild Fire L2
Smoke-Net [13]	2016-2020	Yes	SA, AS, SIB, AU, NA	Landsat-8	54,270	RGB, SWIR2, TIRS1	Manual, Smoke
1-DCNN [14]	2016-2020	No	AU	PRISMA	259 px	RGB, SWIR2, TIRS1	Manual, Fire/Burned Area
3DCNN [15]	2019-2021	No	AU, US, IT	PRISMA	593 px	1-230	Manual, Fire/Burned Area
SN [16]	2019-2021	Yes	TR	Sentinel-2	21,690	1-12	Manual, Fire
VIB_SD [17]	2010-2020	Yes	AU	Landsat 5.8	1.836	RGB, NIR, SWIR1, SWIR2	Manual, Smoke

datasets. In particular, it is noted that the lack of large-scale datasets and the difficulty in obtaining satellite images with high spatial and temporal resolution limit the evaluation and improve the generalization of these models. Studies using additional spectral bands, such as infrared (IR), have shown improvements in fire detection accuracy. In the case of VI_SD, for example, the addition of the NIR band increased accuracy by 6%, while other combinations of IR bands decreased performance, highlighting the need to carefully select the most suitable spectral bands. However, the problem of spectral similarity between smoke and other aerosols, such as clouds or dust, has not yet been completely resolved, and continues to cause false alarms, although to a lesser extent with the use of combined bands.

The findings of this review present significant implications for research and practice. The effectiveness of CNNs in detecting wildfires suggests that these techniques can be integrated into real-time satellite monitoring systems for early warnings. CNN-based models can also improve accuracy in post-fire damage assessment, which is critical for recovery planning and prevention of future fires. However, to maximize the benefits of these technologies, it is essential to address limitations such as the spatial and temporal resolution of satellites and explore onboard processing solutions, which can reduce detection latency. The implementation of models in hardware accelerators, as demonstrated in [14], is a step forward in this direction. Additionally, multi-class datasets represent a considerable challenge, as they require precise and detailed labeling. To resolve the lack of these sets, it is necessary to explore more efficient labeling methods, such as the large-scale dataset generation algorithm used in [10].

To maximize the benefits of these models, it is essential to address current limitations related to the spatial and temporal resolution of satellites, and explore onboard processing to reduce detection latency. Models such as [14]'s 1-DCNN, implemented on hardware accelerators such as Intel Movidius NCS-2 and Nvidia Jetson Nano, have shown that on-board data processing is a viable possibility and could significantly reduce processing times. answer. As for datasets, multiclass datasets have proven to be considerably more complex to classify than binary datasets, due to the greater precision required in labeling, as seen in Table 3. Datasets such as [15], which uses PRISMA images from multiple geographic regions, have revealed up to a 20% drop in accuracy between training and test sets, underscoring the importance of having datasets larger and more diverse.

Furthermore, the useful life of satellites and their orbits directly limit the amount and coverage of data available for fire detection, a fundamental aspect for the development of more robust and generalizable models.

Compared to previous studies, this review highlights significant progress in the accuracy and robustness of CNN models applied to wildfire detection. While previous research focused on threshold-based methods or algorithms with low generalizability, the studies reviewed here demonstrate that combining multiple spectral bands—including the use of infrared bands and hyperspectral imaging experiments such as PRISMA—together With the use of various advanced CNN features and architectures, it enables more accurate and faster fire detection. These approaches have overcome several of the shortcomings of traditional methods, such as the difficulty in setting appropriate thresholds under changing conditions and the high rate of false alarms that resulted from these approaches. However, significant challenges remain, particularly in terms of data quality and the generalizability of models to different geographic contexts. Studies such as [15] have revealed that models that integrate spectral data and spatial information from neighboring pixels can improve accuracy, although only partially, over varied geographic regions.

Regarding model architectures, although in terrestrial and aerial detection it has been mentioned that it is advisable to try models specifically designed for remote sensing applications, it could also be beneficial to explore architectures developed in other areas, such as medical image processing or computer vision in general, which could provide advanced techniques in terms of precision and sensitivity. Additionally, techniques such as neural architecture search (NAS), ensemble models, or transfer learning could optimize model performance more efficiently than manual optimization. In studies such as [9], where NAS was used to optimize a Faster R-CNN architecture in the detection of different types of smoke, an improvement in accuracy and greater adaptability to different conditions was observed, indicating that this approach could be relevant to improve robustness in forest fire detection.

Despite the progress made, this review also reveals several limitations in using CNN for wildfire detection. One of the main limitations is the need for large labeled datasets to train these models, which is expensive and laborious to build. In several studies, such as [15], the lack of sufficiently diverse data led to a significant drop in model performance, revealing a 20% decrease in accuracy between the test and training set when data was insufficient. Furthermore, the generalizability of the models remains a challenge, especially when applied in different geographic regions or under varying environmental conditions. For example, models developed with data from a specific region, such as those from [3] in southern China, may show decreased performance when applied in scenarios with different vegetation or climatic conditions. Another important limitation is the latency in detection, which may be related to the limited spatial resolution of some satellites and the time required to process and transmit the images to Earth, reducing the effectiveness in early detection of fires in real time.

Additionally, the complexity of remote sensing imagery, which typically covers large areas with lower pixel density compared to smaller-scale imagery, and complex backgrounds with different vegetation types or geographic structures, pose additional challenges for accurate fire classification. Although ground and aerial detection faces the same problem of lack of datasets, the creation of these is usually relatively simpler compared to satellite remote sensing, which requires experts who can interpret the complex spectral signatures captured by the satellite sensors. This advantage in data collection for ground and airborne methods could allow for more rapid progress in these areas, suggesting that it would be beneficial to closely monitor these developments and adapt, where possible, the advances made to satellite detection to improve its performance in practice.

Based on the identified limitations, it is recommended that future research focus on the development of more robust models that can generalize better in different contexts and environmental conditions. This is especially relevant considering previous studies, such as [15] and [14], which highlighted how variability in geographic conditions affects the performance of models trained in specific regions. Additionally, it would be beneficial to explore hybrid approaches that combine CNN with other machine learning or image processing techniques, such as transformer-based learning, to improve accuracy and reduce false alarms in different environments. Future research should also focus on the use of hyperspectral imaging, which could offer significant advantages in the detection of active fires and in the differentiation of complex spectral signatures, since hyperspectral images contain up to 230 bands, as in PRISMA [15] , providing critical spectral details.

Likewise, it is crucial to encourage the creation and access to larger and more diverse datasets, as well as the development of technologies that enable data processing on board satellites, thereby reducing latency and improving real-time response capacity. To solve the problem of missing datasets, it is necessary to explore faster and more accurate labeling methods, such as the one used in [10], which used an algorithm to create a large-scale dataset automatically. This method is promising for multiclass datasets, especially if applied to hyperspectral images, as suggested by [12], achieving performance improvements by incorporating a combination of IR and RGB spectral bands.

Furthermore, domain adaptation could be a viable solution to address the lack of specific datasets, allowing the transfer of knowledge from one domain (e.g., images from one satellite) to another (images from another satellite), thus improving the ability of the model to generalize in different scenarios. Creating datasets with similar characteristics, from satellites such as Landsat-8 and Sentinel-2, along with domain adaptation, could result in more robust and accurate models for fire detection. It would be extremely beneficial for future research to make their datasets public, since, as seen in Table 3, most of these are private. The availability of these ensembles would not only allow for more complete evaluations and comparisons between models, but would also help mitigate the problem of missing data for wildfire detection. Once this obstacle is overcome, research could focus on comparing and optimizing specific models for active detection or remote sensing, seeking the optimal architecture for each type of detection.

The review conducted demonstrates that CNNs have considerable potential in wildfire detection using satellite images, although they still face significant challenges, especially in terms of generalization and data availability. Recent advances, such as the combination of multiple spectral bands (including infrared and hyperspectral bands) and the development of advanced architectures, such as FireCNN and Multiscale-Net, have improved detection accuracy and capability. However, maximizing the impact of these technologies in practical applications requires overcoming current limitations, such as the need for large labeled datasets and detection latency. Future studies should focus on the development of robust and diversified datasets, the use of hybrid approaches that integrate various DL techniques, and the implementation of processing on board satellites. These strategies can improve the accuracy and speed of detection, allowing a more effective and timely response to forest fires and their mitigation. Ultimately, these efforts will solidify the role of artificial intelligence in environmental conservation.

However, to maximize the impact of these technologies in practical applications, it is crucial to address current limitations, such as reliance on large labeled datasets and latency in detection. Future research should focus on creating more robust datasets, developing hybrid approaches that integrate various DL techniques, and deploying models with onboard satellite processing. These strategies can not only improve the accuracy and speed of fire detection, but also enable a more effective and timely response in wildfire management and mitigation.

Early detection of wildfires using satellite imagery and CNN offers an invaluable tool for protecting ecosystems and communities. Although technology has advanced significantly, the real challenge lies in achieving models that are robust, accurate and scalable in real environments. Addressing current limitations and promoting access to open datasets will not only strengthen emergency response capacity, but will also enable the full potential of artificial intelligence in environmental conservation and protection.

5. Conclusion

This work has reviewed the state of the art in forest fire detection using CNN using satellite images. Although CNNs have proven to be powerful tools for fire classification at the pixel and scene level, their effectiveness largely depends on data quality and availability. The integration of multiple spectral bands, including hyperspectral bands, has improved detection accuracy; however, significant challenges remain related to generalizability and latency in real-time detection.

The findings of this review underline the relevance of CNNs not only in academic research, but also in practical applications, such as real-time monitoring, with direct impact on wildfire response capacity. Improvements in precision and speed of detection offer the possibility of transforming prevention and response strategies, promoting the development of policies that promote the use of advanced technologies in the management of natural disasters.

The field of remote sensing fire detection is still emerging and faces important limitations, such as the need for large-scale labeled datasets and the difficulty of generalizing models in diverse geographic contexts. These limitations reflect the need to improve both the quality and availability of data and explore new techniques, such as the use of NAS and Ensemble Models in the optimization of architectures such as U-Net, testing their performance on satellite images of wildfires.

Future research should focus on the development of more robust models that can adapt to diverse contexts, thus improving generalization and reducing false alarms. To achieve this, the creation of larger and more accessible datasets could be explored using automatic labeling algorithms that speed up the manual process. Also, the development of hybrid approaches and the implementation of models onboard satellites are promising areas with the potential to significantly reduce latency and improve real-time detection of fires.

This review offers a comprehensive and critical perspective on the use of satellite data and artificial intelligence techniques for wildfire detection. Current advances in hybrid models, advanced architectures, and the integration of multiple data sources reflect significant progress in the field, although some fundamental issues, such as the lack of labeled data, remain unresolved. Addressing these limitations will be essential to achieving the full potential of fire detection technologies. Ultimately, continued research and commitment to the development of innovative solutions in this area will contribute to more effective protection of ecosystems and a faster response in wildfire management.

References

[1] M. N. Suratman, N. H. A. Hamid, M. D. M. Sabri, M. Kusin, and S. A. K. Yamani, *Changes in Tree Species Distribution Along Altitudinal Gradients of Montane Forests in Malaysia*. Springer International Publishing, 2015, pp. 491–522. [Online]. Available: https://link.springer.com/10.1007/978-3-319-12859-7_19

[2] "Deforestation fronts: Drivers and responses in a changing world." [Online]. Available: https://wwfint.awsassets.panda.org/downloads/deforestation_fronts___drivers_and_responses_in_a_changing_world___full_report_1.pdf

[3] Z. Hong, Z. Tang, H. Pan, Y. Zhang, Z. Zheng, R. Zhou, Z. Ma, Y. Zhang, Y. Han, J. Wang, and S. Yang, "Active fire detection using a novel convolutional neural network based on himawari-8 satellite images," *Frontiers in Environmental Science*, vol. 10, 3 2022.

[4] V. E. Sathishkumar, J. Cho, M. Subramanian, and O. S. Naren, "Forest fire and smoke detection using deep learning-based learning without forgetting," *Fire Ecology*, vol. 19, 12 2023.

[5] M. J. Sousa, A. Moutinho, and M. Almeida, "Wildfire detection using transfer learning on augmented datasets," *Expert Systems with Applications*, vol. 142, p. 112975, 3 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417419306931

[6] Y. Zheng, G. Zhang, S. Tan, Z. Yang, D. Wen, and H. Xiao, "A forest fire smoke detection model combining convolutional neural network and vision transformer," *Frontiers in Forests and Global Change*, vol. 6, 4 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/ffgc.2023.1136969/full

[7] N. S. Basturk, "Forest fire detection in aerial vehicle videos using a deep ensemble neural network model," *Aircraft Engineering and Aerospace Technology*, vol. 95, pp. 1257–1267, 7 2023. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/AEAT-01-2022-0004/full/html

[8] Z. Xue, H. Lin, and F. Wang, "A small target forest fire detection model based on yolov5 improvement," *Forests*, vol. 13, p. 1332, 8 2022. [Online]. Available: https://www.mdpi.com/1999-4907/13/8/1332

[9] D. Q. Tran, M. Park, Y. Jeon, J. Bak, and S. Park, "Forest-fire response system using deep-learning-based approaches with cctv images and weather data," *IEEE Access*, vol. 10, pp. 66 061–66 071, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9801825/

[10] A. Rostami, R. Shah-Hosseini, S. Asgari, A. Zarei, M. Aghdami-Nia, and S. Homayouni, "Active fire detection from landsat-8 imagery using deep multiple kernel learning," *Remote Sensing*, vol. 14, p. 992, 2 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/4/992

[11] Z. Wang, T. Peng, and Z. Lu, "Comparative research on forest fire image segmentation algorithms based on fully convolutional neural networks," *Forests*, vol. 13, p. 1133, 7 2022. [Online]. Available: https://www.mdpi.com/1999-4907/13/7/1133

[12] Y. Kang, E. Jang, J. Im, and C. Kwon, "A deep learning model using geostationary satellite data for forest fire detection with reduced detection latency," *GIScience Remote Sensing*, vol. 59, pp. 2019–2035, 12 2022. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/15481603.2022.2143872

[13] H. Liang, C. Zheng, X. Liu, Y. Tian, J. Zhang, and W. Cui, "Super-resolution reconstruction of remote sensing data based on multiple satellite sources for forest fire smoke segmentation," *Remote Sensing*, vol. 15, p. 4180, 8 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/17/4180

[14] K. Thangavel, D. Spiller, R. Sabatini, S. Amici, S. T. Sasidharan, H. Fayek, and P. Marzocca, "Autonomous satellite wildfire detection using hyperspectral imagery and neural networks: A case study on australian wildfire," *Remote Sensing*, vol. 15, p. 720, 1 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/3/720

[15] D. Spiller, A. Carbone, S. Amici, K. Thangavel, R. Sabatini, and G. Laneve, "Wildfire detection using convolutional neural networks and prisma hyperspectral imagery: A spatial-spectral analysis," *Remote Sensing*, vol. 15, p. 4855, 10 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/19/4855

[16] A. M. Al-Dabbagh and M. Ilyas, "Uni-temporal sentinel-2 imagery for wildfire detection using deep learning semantic segmentation models," *Geomatics, Natural Hazards and Risk*, vol. 14, 12 2023. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/19475705.2023.2196370

[17] L. Zhao, J. Liu, S. Peters, J. Li, S. Oliver, and N. Mueller, "Investigating the impact of using ir bands on early fire smoke detection from landsat imagery with a lightweight cnn model," *Remote Sensing*, vol. 14, p. 3047, 6 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/13/3047

Critical perspective

The eXplainable Artificial Intelligence Paradox in Law: Technological Limits and Legal Transparency

Denise M. Trejo-Moncada
Universidad Nacional Autónoma de México, Facultad de Derecho

Keywords: explainability paradox, explainable artificial intelligence, artificial intelligence in law

PERSPECTIVE

The integration of Artificial Intelligence (AI) into legal systems offers transformative potential, promising enhanced efficiency and predictive accuracy. However, this progress also brings to the spotlight the explainability paradox: the unavoidable trade-off between the accuracy of complex Machine Learning (ML) and Deep Learning (DL) models and their lack of transparency. This paradox challenges foundational legal principles such as fairness, due process, and the right to explanation. While eXplainable AI (XAI) techniques have emerged to address this issue, their post-hoc nature, limited fidelity, and inaccessibility to non-expert stakeholders impede their practical utility in legal contexts. This paper critically reflects on the explainability paradox and its implications for AI-assisted legal decision-making, proposing a balanced framework to reconcile accuracy with transparency. By examining the limitations of current XAI methods and exploring the potential of inherently interpretable models, it highlights pathways to align AI systems with the procedural and ethical standards of the legal domain. These reflections not only address a gap in existing research but also challenge conventional reliance on opaque models, advocating for AI systems that prioritize trust, accountability, and legitimacy. This reflection invites interdisciplinary dialogue and encourages the development of AI tools that integrate technical performance with ethical and societal needs, ensuring the responsible adoption of AI in law.

Introduction

Artificial Intelligence (AI) has become an integral component of critical decision-making systems, transforming industries through its ability to analyze complex data and deliver predictive insights [1]. In fields such as healthcare, finance, and transportation, AI-driven systems have demonstrated exceptional performance, enabling faster and more accurate decisions [2, 3]. The legal sector, traditionally reliant on human expertise and interpretability, is increasingly exploring the adoption of AI technologies, particularly Machine Learning (ML) and Deep Learning (DL), to enhance processes such as case management, risk assessment, and evidence analysis [4, 5]. These models, known for their ability to uncover patterns and predict outcomes with unprecedented precision, promise significant advancements in efficiency and accuracy. However, this progress comes at a cost: the inherent opacity of complex ML and DL models raises critical concerns about explainability, trust, and accountability, particularly in contexts where transparency is a foundational requirement [6, 4].

This rapid adoption of AI in legal contexts introduces a fundamental paradox: while the legal system is built on principles of transparency, accountability, and fairness, modern ML and DL models often operate as black boxes [7, 8]. These models generate predictions and decisions with exceptional accuracy, yet their inner workings remain largely opaque, even to their developers [9, 10]. Legal decisions—whether related to sentencing, risk assessment, or evidence evaluation—demand reasoning that is not only accurate but also explainable and justifiable to all stakeholders [11, 12]. In this setting, the lack of interpretability in AI systems conflicts directly with the legal requirement for clear, understandable decision-making. This tension raises critical questions: Can AI systems truly achieve the level of transparency necessary for legal legitimacy, or does their reliance on complexity undermine trust and fairness? [10, 7]

In response to the opacity of modern AI systems, the emerging field of eXplainable AI (XAI) seeks to address the interpretability challenge by providing insights into how complex ML and DL models generate their outputs. Techniques such as SHAP (Shapley Additive Explanations), LIME (Local

E-mail address: denise.trejo@aaaimx.org

https://doi.org/10.5281/zenodo.14692066

Interpretable Model-agnostic Explanations), and visualization methods like Grad-CAM have been developed to offer post-hoc explanations, enabling a degree of transparency in otherwise black-box systems [13, 14]. These tools attempt to highlight the relationships between inputs and predictions, offering approximations of the model's reasoning process [15, 16]. However, despite these advancements, XAI techniques remain limited: they often provide partial or oversimplified explanations that fail to capture the true internal logic of complex models [17, 18]. Furthermore, these explanations are frequently difficult for non-experts to interpret, raising concerns about their adequacy for legal systems where transparency, clarity, and trust are paramount [13, 18]. This gap highlights the ongoing struggle to align the technical capabilities of XAI with the rigorous standards of transparency required in legal decision-making.

This paper critically reflects on the limitations of XAI within the context of legal applications, where transparency and accountability are non-negotiable principles. While ML and DL models offer unprecedented predictive capabilities, their lack of interpretability creates a significant barrier to their adoption in legal systems. This work argues that the paradox between explainability and performance must be addressed to ensure that AI tools align with legal requirements for transparency, fairness, and trust. Without overcoming this challenge, the deployment of AI in sensitive legal contexts risks undermining the very foundations of legitimacy and justice that the legal system upholds. Moving forward, achieving a balance between model performance and explainability must become a central focus for researchers and practitioners seeking to integrate AI into legal decision-making processes responsibly.

Position

The increasing adoption of AI in legal systems brings to the forefront a critical paradox: the tension between accuracy and explainability in ML and DL models [19, 20]. On one hand, the unparalleled predictive power of complex AI systems enables them to identify patterns and deliver insights that surpass human capabilities [21, 22]. On the other hand, the very complexity that drives their performance renders these models opaque and difficult to interpret, even for their creators [23, 24]. In legal contexts, where decisions must be transparent, justifiable, and open to scrutiny, this lack of interpretability presents a significant barrier [25]. However, it is worth noting that human decision-making in legal processes is not without flaws, as biases, inconsistencies, or even corruption can occasionally compromise fairness and accountability. AI systems, despite their opacity, offer the potential for greater objectivity and the ability to inspect and analyze decision-making processes retrospectively. The challenge lies in reconciling the need for high-performing AI systems with the equally important requirement for clear and understandable reasoning—a fundamental pillar of fairness and trust within legal processes [26]. Addressing this paradox is not merely a technical necessity but a critical step toward ensuring the ethical and legitimate integration of AI into the legal domain.

At the core of the explainability paradox is a well-recognized trade-off in AI: simpler, interpretable models often sacrifice accuracy, while complex, high-performing models, such as deep neural networks, sacrifice transparency [27, 28]. Interpretable models—such as decision trees, linear regression, and rule-based systems—are inherently easier to understand and explain [29, 30]. These models allow stakeholders to trace decisions back to specific inputs, offering a level of clarity that aligns with the legal system's need for justification and accountability [31, 32]. However, their simplicity limits their ability to capture intricate patterns within large, complex datasets, often resulting in lower predictive accuracy [33]. Case-Based Reasoning (CBR) offers an alternative approach to bridging this gap by retrieving and reasoning through similar past cases [34]. Its interpretable framework ensures that decisions are supported by precedent [35], which aligns naturally with the legal system's emphasis on contextual and historical consistency.

In contrast, DL models excel at delivering superior performance by leveraging massive datasets and complex architectures to uncover subtle relationships that are imperceptible to traditional approaches [33, 36]. Yet, this complexity comes at the cost of interpretability: the reasoning behind a model's prediction remains opaque, hidden within layers of parameters and computations that defy human understanding [37, 38]. While CBR systems may not achieve the predictive power of DL models, they can enhance interpretability by explicitly tying new decisions to past examples, providing a transparent foundation for explanations. However, it is important to recognize that even opaque DL models may offer opportunities for retrospective inspection and analysis, capabilities that human decision-making processes often lack. In legal applications—where decisions affect lives, freedoms, and rights—accuracy alone is insufficient. The law demands not only correct outcomes but also explanations that can be understood, challenged, and justified [27, 39]. While the opacity of AI systems raises concerns, they may still serve as valuable tools to complement human decision-making, particularly in contexts where human biases or a lack of accountability have historically undermined trust. This trade-off highlights a fundamental challenge in aligning AI's technical capabilities with the ethical and procedural standards required in legal systems [40].

In response to the opacity of complex ML and DL models, significant efforts have been made in developing XAI techniques to make their decision-making processes more interpretable. Methods such as SHAP, LIME, Feature Importance (FI), and visualization tools like Grad-CAM have emerged as widely adopted solutions [13, 18, 16]. These techniques attempt to provide insights into how input features influence model outputs, offering a degree of transparency that was previously unattainable [41, 42]. However, despite these advancements, current XAI methods remain limited in key ways that significantly impact their suitability for legal applications [43, 44].

First, most XAI techniques are post-hoc in nature, meaning they generate explanations after the model has produced its output, rather than ensuring inherent transparency within the model itself [45, 46]. This creates a reliance on approximations, which may not accurately reflect the true reasoning process of the underlying model [47, 48]. Second, the explanations provided by XAI methods are often oversimplified, reducing complex interactions into digestible summaries that can be incomplete or even misleading [49, 8]. For instance, FI scores or Grad-CAM's heatmaps may highlight correlations but fail to convey the relationships or causal factors driving predictions [50]. Lastly, the accessibility of these explanations remains a significant barrier. While XAI outputs may be interpretable for AI experts, they are often too technical or abstract for non-expert stakeholders, such as judges, lawyers, or defendants [51, 52]. In a legal setting—where transparency must be both accurate and comprehensible—these limitations hinder the practical usability of XAI, undermining its ability to meet the rigorous standards of justification and accountability required by the law.

The use of black-box AI systems in legal contexts raises profound implications for fundamental legal principles such as due process, fairness, and the right to explanation [11, 53]. At the heart of the legal system lies the requirement for decisions to be transparent, justifiable, and open to scrutiny [54, 55]. When decisions are influenced or determined by opaque AI models—whose reasoning cannot be fully understood or explained—it becomes challenging for stakeholders to evaluate whether those decisions are fair, unbiased, or free from error [10]. However, it is also worth noting that human decision-makers, despite their interpretability, can at times be equally opaque—whether due to cognitive limitations, unconscious biases, or intentional withholding of reasoning. This parallel suggests that while black-box AI systems introduce challenges, they also provide an opportunity for systematic inspection and reproducibility that human decisions often lack. By documenting decision-making processes through algorithms, AI can offer a framework for identifying errors or biases retrospectively, fostering a level of accountability that is not always achievable in human-driven systems. Addressing the transparency issues inherent in AI is crucial, but leveraging these tools to complement human decision-making could mitigate longstanding concerns about fairness and consistency in the legal domain [56].

Furthermore, the deployment of black-box systems threatens to erode trust—not only in the AI tools themselves but also in the legal institutions adopting them. Trust in the justice system is built on its ability to deliver outcomes that are not only accurate but also explainable and consistent. Opaque AI models, by their nature, introduce uncertainty and raise doubts about accountability, especially when errors occur or biases emerge. However, trust in human decision-making is not always guaranteed either, as it can be compromised by biases, inconsistencies, or even intentional misconduct. AI systems, despite their opacity, offer unique opportunities for post-hoc analysis and continuous improvement, allowing stakeholders to retrospectively evaluate and refine decision-making processes in ways that are often impossible with human decisions. While such systems risk creating a perception that justice is being "outsourced" to inscrutable algorithms, their ability to document and analyze decisions systematically presents a pathway to enhance transparency and accountability if implemented responsibly. Balancing these perspectives is essential to maintaining the legitimacy of legal processes.

Without addressing the explainability problem, the widespread deployment of black-box AI systems could lead to significant ethical and legal challenges, including violations of fairness, potential misuse, and unintended harm. These challenges expose the urgent need for a careful, measured approach to integrating AI in legal decision-making—one that prioritizes transparency and accountability over blind reliance on performance. Failure to resolve these issues may not only weaken confidence in AI but also jeopardize the integrity of the legal system itself.

To bridge the gap between AI's predictive capabilities and the legal system's demand for transparency, it is evident that current solutions remain insufficient. While hybrid approaches and inherently interpretable AI models hold promise, significant technical and ethical challenges persist. Achieving a balance between accuracy and transparency will require ongoing research, responsible design, and a commitment to maintaining human oversight in AI-assisted legal processes. However, it is equally important to recognize AI's potential to complement the legal system by offering consistency and reducing biases that can affect human decision-making. By leveraging AI's ability to document and standardize decision-making processes, legal institutions have an opportunity to evolve toward more equitable and objective outcomes. Until such advancements are realized, the widespread deployment of opaque systems in legal contexts risks undermining trust, accountability, and the very principles the legal system seeks to uphold.

Discussion

The explainability paradox—where increased accuracy in AI systems comes at the cost of transparency—has far-reaching implications for legal systems attempting to integrate ML and DL models. At its core, this paradox challenges the foundational principles of the law, which rely on decisions that are transparent, justifiable, and contestable [57, 58]. Legal systems are not merely tasked with reaching accurate outcomes but must also ensure that decisions can be understood and trusted by all stakeholders, including judges, lawyers, defendants, and the public [59, 60]. While the opacity of AI models can undermine these key pillars of justice, it is essential to recognize that human decision-making is not inherently more transparent. Cognitive biases, inconsistencies, and even deliberate misconduct can obscure the reasoning behind human judgments. In contrast, AI systems provide systematic documentation and opportunities for retrospective analysis, offering a mechanism for identifying errors or biases that might otherwise remain hidden. By leveraging these strengths, AI has the potential to address some of the long-standing challenges associated with human decision-making while complementing traditional legal frameworks.

Moreover, the paradox exposes a deeper ethical tension: efficiency versus legitimacy. While AI systems promise increased efficiency by automating processes and enhancing predictive accuracy, their opacity risks eroding public trust in legal institutions [61, 62]. For example, AI-based risk assessment tools used in sentencing or parole decisions may produce accurate predictions, but without clear explanations, stakeholders cannot assess whether these decisions are fair or free from bias [63]. However, human decision-making is not immune to similar challenges; biases, inconsistencies, and even corruption can sometimes compromise fairness and accountability in legal judgments. In this context, AI systems offer a unique opportunity to mitigate such issues by providing systematic, data-driven insights that are less prone to individual bias and more amenable to post-hoc inspection. Nevertheless, the adoption of opaque AI systems—without addressing their explainability limitations—risks creating a perception that justice is being outsourced to inscrutable algorithms, thereby undermining the legitimacy and credibility of legal outcomes [64]. This tension highlights the urgent need for a more responsible approach to integrating AI into the legal system, one that not only prioritizes technical performance but also leverages AI's potential to enhance fairness while upholding the ethical standards upon which the rule of law is built.

Current efforts in XAI, such as SHAP, LIME, and Grad-CAM, have made progress in shedding light on black-box predictions by offering insights into feature importance and decision boundaries. While these techniques improve transparency, they remain insufficient for legal applications, where clarity, reliability, and accessibility are non-negotiable.

A major limitation of XAI lies in its post-hoc nature—generating explanations after predictions are made. These approximations may fail to reflect the model's true reasoning, raising concerns about their fidelity in sensitive legal decisions. Additionally, XAI outputs are often oversimplified, reducing complex relationships to summaries that can obscure critical connections and mislead stakeholders.

Finally, XAI explanations, such as feature scores or heatmaps, are often too technical for non-expert stakeholders like judges or lawyers. Legal applications require explanations that are clear, actionable, and comprehensible to ensure decisions can be scrutinized and justified [65, 52]. These challenges highlight the need for further advancements in inherently interpretable models that align AI systems with the rigorous standards of fairness and accountability required by the legal system [66, 67].

Hybrid models, which combine interpretable algorithms with the predictive power of black-box systems [68], offer a promising middle ground to address the explainability paradox. For example, interpretable models like decision trees or linear regression could handle critical decisions requiring justification, while complex DL models can assist in auxiliary tasks or pre-processing stages. This layered approach maintains transparency where it matters most while leveraging the accuracy of advanced AI systems. However, hybrid solutions are not without limitations. Seamlessly integrating interpretable and black-box components remains a technical challenge, as discrepancies between models could lead to inconsistencies. Additionally, determining which stages require human oversight versus AI-driven decisions introduces added complexity. Despite these challenges, hybrid models represent a feasible pathway for balancing accuracy and transparency, particularly in legal contexts where trust and accountability are paramount.

A more sustainable solution to the explainability paradox lies in developing inherently interpretable AI models—systems designed for transparency from the outset. Unlike post-hoc explanations, these models integrate interpretability into their structure, ensuring that decision-making processes are both clear and traceable [36]. Recent progress in techniques such as Generalized Additive Models (GAMs), explainable neural networks, and sparse models demonstrates that transparency and performance can coexist, particularly in tasks with structured data [69, 70].

Inherently interpretable models—designed for transparency from the outset—offer a promising solution to the explainability paradox. These models integrate interpretability into their structure, ensuring that decision-making processes are clear and traceable [36]. While such models currently lag behind DL systems in handling unstructured, high-dimensional data, they can mitigate many risks associated with black-box AI in legal contexts. By prioritizing models that are understandable by design, stakeholders gain greater confidence in AI outputs, ensuring alignment with legal standards of fairness and accountability. However, even black-box systems have advantages: their systematic documentation of decision-making pathways enables retrospective analysis and error detection, which are often lacking in human-driven processes. Advancing both inherently interpretable models and strategies to enhance the transparency of complex systems is essential for achieving a balance between accuracy and explainability while addressing the limitations of both human and AI decision-making.

While AI systems offer valuable support in legal decision-making, human oversight remains indispensable. Legal processes require contextual understanding, ethical judgment, and adherence to evolving societal standards—capabilities that AI, regardless of its sophistication, cannot fully replicate [71, 72]. However, AI systems, even when opaque, provide a systematic approach to decision-making that can mitigate human biases and inconsistencies. By combining AI's ability to process large datasets and document decision-making pathways with human expertise in applying broader reasoning, stakeholders can ensure that final decisions remain accountable, fair, and aligned with legal principles. This synergy reduces the risks of relying solely on either human or AI decision-making, creating a collaborative framework where AI enhances consistency and trust while humans provide essential oversight.

The collaborative use of AI and human oversight in legal systems has implications beyond the legal domain, offering a roadmap for ethical AI adoption in other high-stakes areas such as healthcare, finance, and governance. Lessons learned from aligning AI with legal standards—such as prioritizing transparency, fairness, and accountability—can inform the responsible deployment of AI across industries. Future research must focus on developing inherently interpretable models that handle unstructured, high-dimensional data without sacrificing performance, as well as refining hybrid approaches to strike a balance between accuracy and transparency. Additionally, interdisciplinary collaboration between AI researchers, legal scholars, and ethicists will be critical to ensuring that AI systems are not only technically robust but also ethically aligned with societal values. By addressing these challenges, AI can evolve into a tool that enhances decision-making processes while maintaining trust and accountability in sensitive applications.

Resolving the explainability paradox has implications beyond the legal domain, offering a roadmap for AI adoption in other high-stakes areas like healthcare, finance, and governance. Lessons learned from aligning AI with legal standards—such as prioritizing transparency, fairness, and accountability—can inform ethical AI practices across industries. Future research must focus on developing inherently interpretable models capable of handling complex, unstructured data without sacrificing performance, as well as refining hybrid approaches to strike a balance between accuracy and explainability.

Additionally, interdisciplinary collaboration between AI researchers, legal scholars, and ethicists will be critical to ensuring AI systems are not only technically robust but also ethically aligned with societal values. By addressing these challenges, AI can evolve into a tool that enhances decision-making processes while maintaining trust and accountability in sensitive applications.

Conclusions

This paper has examined the explainability paradox in AI: the tension between the exceptional predictive performance of ML and DL models and their lack of transparency. While these models hold the potential to revolutionize legal systems, their opacity raises significant challenges for ensuring that decisions remain transparent, justifiable, and accountable. This paradox underscores the critical need to align AI tools with legal principles, ensuring their integration enhances the integrity and fairness of legal processes rather than undermining them.

Current XAI techniques, though instrumental in improving transparency, remain limited by their post-hoc nature, oversimplification of complex model behavior, and inaccessibility to non-experts. These limitations are particularly pronounced in legal contexts, where clarity and justification are paramount. To address this, hybrid models that combine interpretable algorithms with the predictive power of black-box systems offer a practical interim solution. In the longer term, the development of inherently interpretable AI systems, designed for transparency without sacrificing performance, represents a more sustainable path forward. Crucially, maintaining human oversight remains indispensable to uphold fairness, accountability, and public trust.

Moving forward, addressing the explainability paradox requires a collaborative effort between researchers, developers, and legal practitioners. Future work should prioritize advancing inherently interpretable models, refining XAI techniques to enhance fidelity and accessibility, and fostering interdisciplinary collaboration between AI experts, legal scholars, and ethicists. By aligning technological advancements with ethical and legal standards, we can harness AI's potential to complement human judgment, mitigate biases, and strengthen the principles of fairness and trust that underpin the legal system.

CRediT authorship contribution statement

Denise Trejo-Moncada: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used ChatGPT in order to improve readability. After using this tool, the author reviewed and edited the content as needed and took full responsibility for the content of the publication.

Declaration of competing interest

The author declares that she has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author would like to express her deepest gratitude to F.E. Heart for the invaluable discussions that greatly enhanced the depth of this work. Special thanks also go to D.P. Assenger, whose insightful feedback during the revision process was instrumental in refining the manuscript. Their contributions have been truly appreciated.

References

[1] M. Elhaddad and S. Hamam, “Ai-driven clinical decision support systems: An ongoing pursuit of potential,” *Cureus*, vol. 16, 2024.

[2] M. Jeyaraman, S. Balaji, N. Jeyaraman, and S. Yadav, “Unraveling the ethical enigma: Artificial intelligence in healthcare,” *Cureus*, vol. 15, 2023.

[3] K. Yekaterina, “Challenges and opportunities for ai in healthcare,” *International Journal of Law and Policy*, 2024.

[4] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, “Trustworthy artificial intelligence: A review,” *ACM Computing Surveys (CSUR)*, vol. 55, pp. 1–38, 2022.

[5] V. Lai, C. Chen, A. Smith-Renner, Q. Liao, and C. Tan, “Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies,” in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023.

[6] N. Biller-Andorno, A. Ferrario, S. Joebges *et al.*, “Ai support for ethical decision-making around resuscitation: proceed with care,” *Journal of Medical Ethics*, vol. 48, pp. 175–183, 2020.

[7] N. Thalpage, “Unlocking the black box: Explainable artificial intelligence (xai) for trust and transparency in ai systems,” *Journal of Digital Art & Humanities*, 2023.

[8] G. Chaudhary, “Explainable artificial intelligence (xai): Reflections on judicial system,” *Kutafin Law Review*, 2024.

[9] M. S. M. d. Encarnacao, M. Anastasiadou, and V. Santos, “Framework for the application of explainable artificial intelligence techniques in the service of democracy,” *Transforming Government: People, Process and Policy*, 2024.

[10] V. Gupta, S. Shukla, and K. Nikita, “Cracking the code: Enhancing trust in ai through explainable models,” *resmilitaris*, 2024.

[11] M. T. Sacramed, “Reviewing the philippines legal landscape of artificial intelligence (ai) in business: Addressing bias, explainability, and algorithmic accountability,” *International Journal of Research and Innovation in Social Science*, 2024.

[12] M. S. Marques da Encarnacao, M. Anastasiadou, and V. Santos, “Framework for the application of explainable artificial intelligence techniques in the service of democracy,” *Transforming Government: People, Process and Policy*, 2024.

[13] L. Zou, H. Goh, C. Liew, J. Quah, G. T. Gu, J. J. Chew, M. P. Kumar, C. Ang, and A. W. A. Ta, “Ensemble image explainable ai (xai) algorithm for severe community-acquired pneumonia and covid-19 respiratory infections,” *IEEE Transactions on Artificial Intelligence*, vol. 4, pp. 242–254, 2023.

[14] H. Byeon, “Advances in machine learning and explainable artificial intelligence for depression prediction,” *International Journal of Advanced Computer Science and Applications*, 2023.

[15] A. G, S. B. Madagaonkar, and R. C. H, “Unveiling the black box: A comprehensive review of explainable ai techniques,” *International Journal of Scientific Research in Engineering and Management*, 2024.

[16] F. A. Undie, L. V. Kruglova, M. O. Okache, V. A. Undie, and R. A. Aloye, “Exploring explainable artificial intelligence (xai) to enhance healthcare decision support systems in nigeria,” *Journal of Innovative Research*, 2024.

[17] M. Saarela and V. Podgorelec, “Recent applications of explainable ai (xai): A systematic literature review,” *Applied Sciences*, 2024.

[18] F. Abdullakutty, Y. Akbari, S. Al-Maadeed, A. Bouridane, I. M. Talaat, and R. Hamoudi, “Histopathology in focus: a review on explainable multi-modal approaches for breast cancer diagnosis,” *Frontiers in Medicine*, 2024.

[19] S. Veer, L. Riste, S. Cheraghi-Sohi *et al.*, “Trading off accuracy and explainability in ai decision-making: findings from 2 citizens’ juries,” *Journal of the American Medical Informatics Association*, vol. 28, pp. 2128–2138, 2021.

[20] K. Atkinson, T. J. M. Bench-Capon, and D. Bollegala, “Explanation in ai and law: Past, present and future,” *Artificial Intelligence*, vol. 289, p. 103387, 2020.

[21] G. Chaudhary, “Explainable artificial intelligence (xai): Reflections on judicial system,” *Kutafin Law Review*, 2024.

[22] R. Ejjami, “Ai-driven justice: Evaluating the impact of artificial intelligence on legal systems,” *International Journal For Multidisciplinary Research*, 2024.

[23] G. Vilone and L. Longo, “Notions of explainability and evaluation approaches for explainable artificial intelligence,” *Information Fusion*, vol. 76, pp. 89–106, 2021.

[24] W. Guo, “Explainable artificial intelligence for 6g: Improving trust between human and machine,” *IEEE Communications Magazine*, vol. 58, pp. 39–45, 2020.

[25] G. Joshi, “A systematic review on explainable ai in legal domain,” *International Journal for Research in Applied Science and Engineering Technology*, 2024.

[26] T. Ha, S. Lee, and S. Kim, “Designing explainability of an artificial intelligence system,” in *Proceedings of the Technology, Mind, and Society*, 2018.

[27] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, “Definitions, methods, and applications in interpretable machine learning,” *Proceedings of the National Academy of Sciences*, vol. 116, pp. 22 071–22 080, 2019.

[28] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable ai: A review of machine learning interpretability methods,” *Entropy*, vol. 23, 2020.

[29] M. Nauta, J. Trienes, S. Pathak *et al.*, “From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai,” *ACM Computing Surveys*, 2022.

[30] C. P. R. Vieira and L. A. Digiampietri, “Machine learning post-hoc interpretability: A systematic mapping study,” *XVIII Brazilian Symposium on Information Systems*, 2022.

[31] M. M. Xu, P. Watkinson, and T. Zhu, “Explainable ai for clinical risk prediction: A survey of concepts, methods, and modalities,” *ArXiv*, vol. abs/2308.08407, 2023.

[32] K. Sankaran, “Data science principles for interpretable and explainable ai,” *ArXiv*, vol. abs/2405.10552, 2024.

[33] Q. Zhang and S.-C. Zhu, “Visual interpretability for deep learning: a survey,” *Frontiers of Information Technology & Electronic Engineering*, vol. 19, pp. 27–39, 2018.

[34] J. A. Recio-García, H. Parejas-Llanovarced, M. G. Orozco-del Castillo, and E. E. Brito-Borges, “A case-based approach for the selection of explanation algorithms in image classification,” in *Case-Based Reasoning Research and Development: 29th International Conference, ICCBR 2021, Salamanca, Spain, September 13–16, 2021, Proceedings 29*. Springer, 2021, pp. 186–200.

[35] M. G. Orozco-del Castillo, J. A. Recio-Garcia, and E. C. Orozco-del Castillo, “Item-specific similarity assessments for explainable depression screening,” in *International Conference on Case-Based Reasoning*. Springer, 2024, pp. 430–444.

[36] A. Somani, A. Horsch, A. Bopardikar, and D. K. Prasad, “Propagating transparency: A deep dive into the interpretability of neural networks,” *Nordic Machine Intelligence*, 2024.

[37] X. Li, H. Xiong *et al.*, “Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond,” *Knowledge and Information Systems*, vol. 64, pp. 3197–3234, 2021.

[38] L. Munroe, M. d. Silva *et al.*, “Applications of interpretable deep learning in neuroimaging: A comprehensive review,” *Imaging Neuroscience*, vol. 2, pp. 1–37, 2024.

[39] A. M. Hanif, S. Beqiri, P. Keane, and J. Campbell, “Applications of interpretability in deep learning models for ophthalmology,” *Current Opinion in Ophthalmology*, vol. 32, pp. 452–458, 2021.

[40] F. Fan, J. Xiong, M. Li, and G. Wang, “On interpretability of artificial neural networks: A survey,” *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, pp. 741–760, 2020.

[41] A. G, S. B. Madagaonkar, and R. C. H, “Unveiling the black box: A comprehensive review of explainable ai techniques,” *International Journal of Scientific Research in Engineering and Management*, 2024.

[42] D. Tang, J. Chen, L. Ren, X. Wang, D. Li, and H. Zhang, “Reviewing cam-based deep explainable methods in healthcare,” *Applied Sciences*, 2024.

[43] N. Y. Fares, D. Nedeljkovic, and M. Jammal, “Ai-enabled iot applications: Towards a transparent governance framework,” *2023 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, pp. 109–114, 2023.

[44] F. Xu, L. Jiang, W. He, G. Huang, Y. Hong, F. Tang, J. Lv, Y. Lin, Y. Qin, R. Lan, X. Pan, S. Zeng, M. Li, Q. Chen, and N. Tang, “The clinical value of explainable deep learning for diagnosing fungal keratitis using in vivo confocal microscopy images,” *Frontiers in Medicine*, vol. 8, 2021.

[45] A. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney, “Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review,” *Applied Sciences*, vol. 11, p. 5088, 2021.

[46] B. M. de Vries, G. Zwezerijnen, G. Burchell, F. V. van Velden, C. W. M. van der Houven van Oordt, and R. Boellaard, “Explainable artificial intelligence (xai) in radiology and nuclear medicine: a literature review,” *Frontiers in Medicine*, vol. 10, 2023.

[47] M. Mainali and R. O. Weber, “What’s meant by explainable model: A scoping review,” *ArXiv*, vol. abs/2307.09673, 2023.

[48] C. P. R. Vieira and L. A. Digiampietri, "Machine learning post-hoc interpretability: a systematic mapping study," in *XVIII Brazilian Symposium on Information Systems*, 2022.

[49] M. Fontes, J. D. S. D. Almeida, and A. Cunha, "Application of example-based explainable artificial intelligence (xai) for analysis and interpretation of medical imaging: A systematic review," *IEEE Access*, vol. 12, pp. 26 419–26 427, 2024.

[50] M. T. Keane and E. M. Kenny, "How case-based reasoning explains neural networks: A theoretical analysis of xai using post-hoc explanation-by-example from a survey of ann-cbr twin-systems," in *Lecture Notes in Computer Science*, 2019, pp. 155–171.

[51] M. Velmurugan, C. Ouyang, Y. Xu, R. Sindhgatta, B. Wickramanayake, and C. Moreira, "Developing guidelines for functionally-grounded evaluation of explainable artificial intelligence using tabular data," in *2024 International Conference on Explainable AI*, 2024.

[52] Z. Zhou, M. Hu, M. Salcedo, N. Gravel, W. Yeung, A. Venkat, D. Guo, J. Zhang, N. Kannan, and S. Li, "Xai meets biology: A comprehensive review of explainable ai in bioinformatics applications," *ArXiv*, vol. abs/2312.06082, 2023.

[53] P. N. Thakre, P. R. Sahu, P. K. Soni, D. Bisen, and A. Sahu, "Evaluating transparency and explainability in ai-driven planning and scheduling: A comprehensive literature review," *International Journal of Innovative Research in Science, Engineering and Technology*, 2023.

[54] M. I. Konkov, "Ethical issues of implementing artificial intelligence in medicine," *Digital Diagnostics*, 2023.

[55] G. Chaudhary, "Explainable artificial intelligence (xai): Reflections on judicial system," *Kutafin Law Review*, 2024.

[56] K. Ingram, "Ai and ethics: Shedding light on the black box," *International Review of Information Ethics*, 2020.

[57] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, and J. Qadir, "Explainable, trustworthy, and ethical machine learning for healthcare: A survey," *Computers in biology and medicine*, vol. 149, p. 106043, 2021.

[58] A. Kale, T. C. Nguyen, F. Harris, C. Li, J. Zhang, and X. Ma, "Provenance documentation to enable explainable and trustworthy ai: A literature review," *Data Intelligence*, vol. 5, pp. 139–162, 2022.

[59] U. Bhatt, Y. Zhang, J. Antorán, Q. Liao, P. Sattigeri, R. Fogliato, G. G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, L. Nachman, R. Chunara, A. Weller, and A. Xiang, "Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty," *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2020.

[60] M. Hossain, S. Das, B. Krishnamurthy, and S. G. Shiva, "Explainability of artificial intelligence systems: A survey," *2023 International Symposium on Networks, Computers and Communications (ISNCC)*, pp. 1–6, 2023.

[61] B. C. Cheong, "Transparency and accountability in ai systems: safeguarding wellbeing in the age of algorithmic decision-making," *Frontiers in Human Dynamics*, 2024.

[62] G. B. Mensah, "Artificial intelligence and ethics: A comprehensive reviews of bias mitigation,transparency, and accountability in ai systems," *Africa Journal For Regulatory Affairs*, 2024.

[63] E. E. Agu, A. O. Abhulimen, A. N. Obiki-Osafiele, O. S. Osundare, I. A. Adeniran, and C. P. Efunniyi, "Discussing ethical considerations and solutions for ensuring fairness in ai-driven financial services," *International Journal of Frontline Research in Multidisciplinary Studies*, 2024.

[64] S. Akter, "Ethical ai development for sustainable enterprises: A review of integrating responsible ai with iot and enterprise systems," *Journal of Artificial Intelligence General Science*, 2024.

[65] E. Owens, B. Sheehan, M. Mullins, M. Cunneen, J. Ressel, and G. Castignani, "Explainable artificial intelligence (xai) in insurance," *Risks*, 2022.

[66] S. Alam and Z. Altıparmak, "Xai-cf - examining the role of explainable artificial intelligence in cyber forensics," *ArXiv*, vol. abs/2402.02452, 2024.

[67] A. Kalyakulina and I. Yusipov, "explainable artificial intelligence (xai) in age prediction: A systematic review," *ArXiv*, vol. abs/2307.13704, 2023.

[68] G. Q. Álvarez, M. J. del Jesús Díaz, and P. G. García, "Explainable artificial intelligence: An overview on hybrid models," in *Proceedings of the First Multimodal, Affective and Interactive eXplainable AI Workshop (MAI-XAI24 2024), co-located with 27th European Conference on Artificial Intelligence (ECAI 2024)*, J. M. Alonso-Moral, Z. Anthis, R. Berlanga, A. Catalá, P. Cimiano, P. Flach, E. Hüllermeier, T. Miller, O. Mitruţ, D. Mindlin, G. Moise, A. Moldoveanu, F. Moldoveanu, K. Sokol, and A. Soroa, Eds. Santiago de Compostela, Spain: CEUR Workshop Proceedings, 2024, pp. 49–60. [Online]. Available: http://ceur-ws.org/Vol-3803/

[69] C. Lee, M. Samad, I. Hofer, M. Cannesson, and P. Baldi, "Development and validation of an interpretable neural network for prediction of postoperative in-hospital mortality," *NPJ Digital Medicine*, vol. 4, 2021.

[70] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, 2020.

[71] M. Yurrita, T. Draws, A. Balayn, D. Murray-Rust, N. Tintarev, and A. Bozzon, “Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[72] K. S. Reddy, M. Kethan, S. M. Basha, A. Singh, P. Kumar, and D. Ashalatha, “Ethical and legal implications of ai on business and employment: Privacy, bias, and accountability,” in *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, vol. 1, 2024, pp. 1–6.

Journal of Artificial Intelligence and Computing Applications (2024) 2(1): 28–34

Critical perspective

Empathy, Accessibility, and Transparency in the Future of Artificial Intelligence: A Critical Perspective on Technology's Role in Modern Life

Isabel del-Castillo-Borja
International Institute for Intelligent Technologies

Keywords: ethical AI development, AI and social impact, human-centric technology

PERSPECTIVE

Artificial intelligence (AI) has profoundly reshaped daily life, offering unprecedented efficiency and convenience across sectors such as healthcare, education, and agriculture. However, its rapid adoption raises critical concerns about its unintended societal consequences. This article argues that while AI enhances human capabilities, overreliance on its automation risks eroding creativity, critical thinking, and interpersonal connections, particularly among younger generations. Building on existing research, this paper adopts a critical perspective to highlight the dual nature of AI: its potential to empower individuals and address complex challenges versus its propensity to displace cultural practices, weaken community bonds, and foster passivity. Using real-world examples and an interdisciplinary approach, we position AI as a tool that must align with core human values such as empathy, accessibility, and transparency. We advocate for a reimagining of AI's role as a supportive partner rather than a replacement for human agency. From the position that AI must align with ethical principles, this paper argues that fostering user understanding of its limitations and prioritizing initiatives that promote human connection can enable AI to serve as a catalyst for innovation without undermining the social fabric. This article presents a perspective that remarks the urgent need for thoughtful AI development to ensure technology complements humanity rather than diminishes it.

Introduction

Artificial intelligence (AI) has clearly become a transformative force, reshaping the way individuals interact with technology and the world around them. From simplifying mundane tasks [1, 2] to providing instant access to information [3], AI systems have proven indispensable in enhancing efficiency and convenience. Yet, this rapid integration raises critical questions about the long-term implications for creativity, empathy, and social connectedness. As AI evolves, it is imperative to reflect on its impact beyond functionality, addressing how it aligns with fundamental human values.

Despite its many benefits, AI's growing presence has sparked concerns about its unintended consequences [4]. By automating problem-solving and simplifying tasks, AI may risk fostering apathy and reducing opportunities for creativity and critical thinking [5]. Moreover, younger generations may prioritize technological convenience over meaningful human interaction [6], potentially weakening the social fabric that underpins personal and communal growth [7]. These issues highlight the need for a more thoughtful approach to AI development—one that considers its societal and cultural implications.

This paper adopts a critical perspective on AI's role in modern life, emphasizing the importance of aligning its development with principles of empathy, accessibility, and transparency. By examining both the opportunities and challenges posed by AI, the paper argues for a balanced approach that enhances human life while preserving essential values. The discussion highlights the need for AI systems to not only serve practical purposes but also foster trust, inclusivity, and human connection.

Through this lens, the paper aims to contribute to the broader discourse on AI, offering actionable insights for researchers, developers, and policymakers. It advocates for AI technologies that empower individuals without compromising creativity, interaction, or imagination. By prioritizing these considerations, AI can evolve into a tool that truly enriches lives, bridging the gap between technological advancement and human-centric design.

E-mail address: isabel.del-castillo@iiitechs.org
https://doi.org/10.5281/zenodo.14693083

Position

AI has demonstrated its potential to revolutionize daily life, from streamlining mundane tasks to expanding access to knowledge. However, this transformative power comes with significant challenges that warrant critical reflection. This section presents a critical position on the role of AI, focusing on three key areas: its positive impacts, the challenges and concerns it raises, and the core values that should guide its development. By examining these dimensions, the paper articulates a vision for AI that balances innovation with a commitment to human wellbeing, creativity, and social connection.

AI has emerged as a transformative force, fundamentally altering how individuals approach daily tasks and access information. By automating repetitive processes and streamlining complex workflows [8], AI has significantly enhanced efficiency across a range of applications, from personal [9] to professional [10] contexts. Its capacity to provide quick and accurate solutions has not only saved users considerable time but has also minimized the need for extensive manual effort [11]. These advancements have empowered individuals to focus on higher-value activities while fostering a more productive and interconnected society [12].

The real-world applications of AI evidence its profound impact on daily life, offering practical solutions to everyday challenges. Virtual assistants, for instance, have revolutionized how individuals access critical information, such as medical guidance [13]. With a simple voice command, common users can, for instance, obtain details about medications, side effects, and contraindications—tasks that once required consulting outdated books or unreliable sources [14]. Similarly, AI-powered streaming platforms like Spotify have transformed entertainment, curating personalized playlists that eliminate the need for manual selection while enhancing user satisfaction [15]. Beyond these conveniences, AI also supports creative endeavors, such as automating mundane kitchen tasks to allow users to focus on cooking per se [16]. These applications illustrate AI's versatility and its ability to integrate seamlessly into various facets of life, enriching both routine and recreational activities.

By automating mundane or resource-intensive tasks, AI empowers individuals to redirect their time and mental effort toward more meaningful and creative pursuits. Tasks that once demanded significant manual labor, such as managing schedules, organizing information, or performing repetitive calculations, are now streamlined through AI-powered tools [8]. This shift not only reduces cognitive overload but also fosters an environment where innovation and personal growth can flourish [17]. For instance, AI-driven tools like automated transcription services save time by converting meeting recordings into detailed, searchable text, allowing individuals to focus on analyzing and implementing decisions [18].

AI has significantly enhanced accessibility and inclusivity by breaking down barriers to resources and opportunities. For individuals who face challenges such as outdated information sources or limited mobility, AI-powered tools provide a lifeline [19]. Virtual assistants, as aforementioned, deliver real-time answers to queries that would otherwise require navigating complex or unreliable sources [20]. Similarly, AI technologies have transformed access to education and professional development by offering adaptive learning platforms and personalized resources tailored to individual needs [21]. Beyond information, AI-driven innovations in mobility solutions—such as navigation aids for visually impaired users—exemplify its potential to empower marginalized groups [22]. By bridging gaps in knowledge and opportunity, AI holds the promise of fostering a more equitable and inclusive society, ensuring that its benefits are accessible to all.

However, these advancements also come with unintended consequences that warrant critical reflection. The increasing reliance on AI for problem-solving, while convenient, raises concerns about its impact on human creativity and critical thinking [5]. As AI systems provide instant answers and solutions, individuals may become less inclined to engage in independent thought or explore alternative approaches to challenges [23]. This dependency risks fostering a passive mindset, where convenience outweighs the effort of innovation. Over time, the ease of access to AI-driven solutions could erode the skills needed to think critically and creatively, stifling the ability to generate unique ideas or tackle complex problems without technological aid [24]. Such a shift highlights the importance of balancing AI integration with opportunities for individuals to develop and exercise their creative potential.

The pervasive use of AI and digital technologies has contributed to a noticeable decline in face-to-face communication and shared experiences [25], particularly among younger generations [26]. As individuals increasingly rely on virtual interactions and AI-driven platforms for communication and entertainment, opportunities for genuine human connection are diminished [27]. This shift may not only weaken empathy but also undermine the foundations of community building and personal development. For example, activities that once fostered collaboration and mutual understanding, such as group discussions or shared recreational experiences, are now often replaced by solitary engagement with technology [28]. Over time, this trend risks creating a more fragmented society, where meaningful interpersonal relationships and the social skills necessary for cooperative growth are deprioritized.

Despite its advancements, AI often falls short of meeting user expectations due to limitations in current technologies. Frequent errors in tools like voice recognition systems [29], for instance, can lead to frustration and mistrust among users [30]. These missteps highlight a significant gap between the seamless, intuitive performance users expect and the reality of AI's capabilities. Inaccurate responses, misinterpretations, and system glitches undermine the reliability of AI, particularly in contexts where precision is critical, such as healthcare or accessibility tools [31].

To address these challenges and foster trust, inclusivity, and reliability, AI development must be guided by three core values: empathy, accessibility, and transparency. These principles not only ensure that AI systems meet the practical needs of diverse users but also uphold ethical standards that prioritize human well-being. By embedding empathy into AI designs, enhancing accessibility for all, and maintaining transparency in decision-making processes, developers can create technologies that empower users while bridging gaps in trust and functionality.

Embedding empathy into AI systems is crucial for ensuring they effectively address human needs. In fields such as healthcare, AI has the potential to compensate for the impersonal or inadequate experiences often encountered in traditional services. For instance, empathetic AI diagnostic tools could provide patients with detailed explanations of their conditions, treatment options, and potential side effects, fostering a sense of understanding and care [32]. By tailoring responses to individual needs and preferences, AI systems can alleviate feelings of neglect or frustration that arise from hurried consultations with overburdened professionals [33]. Such empathetic designs not only enhance user satisfaction but also establish AI as a trustworthy and supportive tool in addressing complex human challenges [34].

Designing AI technologies with accessibility in mind is essential to ensure they serve a diverse range of users, regardless of age, technical proficiency, or physical ability [35]. Inclusive AI systems can bridge gaps in opportunity by providing tailored solutions for those who face unique challenges. For example, AI-powered tools such as screen readers or voice-controlled assistants have transformed accessibility for individuals with visual or motor impairments, enabling greater independence in daily activities [36]. Similarly, adaptive learning platforms can customize educational content to meet the needs of users with varying levels of digital literacy or cognitive abilities. By prioritizing inclusivity, AI technologies can foster equity and empower marginalized communities, ensuring their benefits are widely distributed.

Transparency plays a pivotal role in building trust and reliability in AI systems [37]. Users need to understand how AI makes decisions and the limitations of its capabilities to engage confidently and responsibly with these technologies. For instance, providing clear explanations of an AI's reasoning process—such as how it arrived at a medical diagnosis or recommended a specific course of action—can help users make informed decisions [38]. Similarly, openly communicating potential biases or areas of uncertainty within an AI system can prevent overreliance and mitigate risks [39]. By fostering openness, transparency not only strengthens user trust but also encourages developers to uphold ethical standards in AI design and deployment.

Discussion

One of the most significant advantages of AI lies in its ability to automate mundane and repetitive tasks, allowing individuals to focus their time and energy on more meaningful and creative endeavors. From scheduling appointments to managing data entry, AI streamlines processes that once required substantial manual effort. For example, virtual assistants like Alexa or Google Assistant can handle everyday queries and reminders, enabling users to redirect their attention toward strategic decision-making or personal pursuits. Similarly, AI tools in industries such as manufacturing or logistics optimize workflows, boosting productivity and efficiency [40]. By offloading routine responsibilities, AI not only enhances convenience but also creates opportunities for individuals to engage in activities that foster innovation and personal growth.

While the automation of routine tasks offers undeniable benefits, over-reliance on AI poses significant risks, particularly to critical thinking and problem-solving skills [41]. As users increasingly depend on AI systems for instant answers, such as relying on virtual assistants to resolve queries, they may become less inclined to engage in independent analysis or explore creative solutions. This dependency fosters a passive mindset, where the convenience of pre-packaged solutions discourages the effort required for deeper cognitive engagement [42]. For example, students turning to AI-powered tools for quick homework answers may bypass the learning process entirely, missing opportunities to develop analytical skills [43]. Such over-reliance not only diminishes individual capabilities but also risks stifling innovation at a broader societal level, evidencing the need for a balanced integration of AI into daily life.

AI holds the potential to serve as a complementary tool to human effort, augmenting capabilities rather than replacing them. In healthcare, for example, AI can enhance diagnostics by analyzing vast datasets with precision, identifying patterns that may be missed by human practitioners [44]. This technological support allows medical professionals to make more accurate and timely decisions, ultimately improving patient outcomes [45]. However, AI's role should remain supportive, preserving the irreplaceable value of empathetic human interactions in patient care. A compassionate practitioner not only interprets data but also understands the emotional and social dimensions of a patient's experience—an aspect that AI, despite its computational power, cannot (yet) replicate. By combining AI's efficiency with human empathy, healthcare can achieve a balance that leverages the strengths of both.

Beyond healthcare, the supportive role of AI extends to a range of fields, amplifying human capabilities while respecting the value of personal judgment and expertise. In education, AI-powered platforms can customize learning experiences to meet diverse student needs, enabling teachers to focus on fostering creativity and critical thinking [46]. In agriculture, AI systems optimize resource management and crop monitoring, equipping farmers with data-driven insights to improve yield and sustainability [47]. Similarly, in disaster response, AI facilitates rapid analysis of satellite imagery and real-time communication, enabling responders to act with greater efficiency and precision [48]. These applications demonstrate how AI can enhance human effectiveness across sectors, driving progress while maintaining the integrity of human decision-making.

The rapid adoption of AI and related technologies poses a significant risk to cultural and social practices, particularly among younger generations. As technology increasingly mediates communication and entertainment, traditional forms of interaction and shared experiences are often displaced [49]. For example, social gatherings that once revolved around communal activities, such as storytelling, board games, or shared meals, are now frequently replaced by isolated interactions with AI-driven devices or online platforms [50]. This shift reduces opportunities for face-to-face engagement, weakening the bonds that foster empathy and community cohesion. Additionally, cultural practices that rely on personal connection and transmission—such as oral traditions or hands-on mentorship—may struggle to adapt in a technology-dominated environment, further emphasizing the need to balance innovation with the preservation of social and cultural heritage [51].

The decline in interpersonal interaction driven by overreliance on AI and digital technologies may carry profound long-term societal consequences [52]. Reduced face-to-face communication erodes empathy [53], a critical skill for understanding and relating to others. As individuals become increasingly detached from direct human connection, the foundations of community bonds weaken, potentially leading to more fragmented and isolated societies. This disconnection undermines collective resilience and the shared values that enable communities to thrive. To mitigate these risks, it is essential to balance technological advancements with efforts to preserve and promote human connection. By prioritizing initiatives that encourage collaborative activities and in-person engagement, society can ensure that technology complements, rather than replaces, the rich social fabric that underpins collective well-being.

Developing AI systems with a human-centric and sustainable approach is essential to maximize their utility and long-term impact. By actively incorporating user feedback during the design process, developers can create technologies that prioritize usability and empathy, ensuring they address real-world needs effectively. For example, tailoring interfaces to accommodate diverse user groups—such as older adults or individuals with disabilities—enhances accessibility and fosters trust in AI solutions. In parallel, sustainability in AI design must be emphasized by promoting long-term reusability and reducing waste. This includes creating modular systems that can be upgraded or repurposed, minimizing the need for constant replacement and reducing the environmental footprint of AI technologies.

The establishment of robust ethical frameworks is critical to ensuring that AI aligns with societal values and promotes equitable progress. Policymakers, researchers, and developers must collaborate to define clear guidelines that address AI's impact on creativity, social interaction, and accessibility. For instance, ethical standards could mandate that AI systems are designed to complement, rather than undermine, human creativity by encouraging collaborative processes instead of replacing them. Similarly, regulations should prioritize technologies that foster social connection rather than perpetuate isolation, particularly among vulnerable populations. Ensuring accessibility must also remain a cornerstone of these frameworks, guaranteeing that AI benefits are distributed broadly across all demographics. Through interdisciplinary collaboration and a shared commitment to ethical principles, stakeholders can develop AI systems that advance innovation while upholding the values of empathy, accessibility, and transparency .

Conclusions

AI holds remarkable potential to transform daily life, from automating mundane tasks to enhancing productivity and supporting human efforts across diverse sectors. By streamlining workflows and providing data-driven insights, AI empowers individuals and organizations to achieve greater efficiency and innovation. However, alongside these benefits come significant challenges. Over-reliance on AI risks diminishing creativity, critical thinking, and problem-solving skills, while excessive use of technology may erode social interaction and cultural practices. These dualities calls for integrating AI into daily life in a balanced manner, ensuring it complements human strengths rather than undermining them.

To harness the full potential of AI while mitigating its challenges, it is essential to embed core values such as empathy, accessibility, and transparency into its development. AI systems must prioritize understanding and addressing human needs, ensuring they remain inclusive and user-friendly for individuals of all backgrounds and abilities. Equally critical is the transparent design of AI, fostering trust and empowering users to engage responsibly with these technologies. However, without these guiding principles, AI risks exacerbating societal issues, including cultural displacement and the erosion of community bonds. As technology continues to evolve, preserving human connection and fostering creativity must remain central goals, ensuring that AI enhances rather than diminishes the richness of interpersonal and cultural experiences.

The future of AI depends on a collective commitment to adopting human-centric, sustainable, and ethically sound approaches to its development. Policymakers, researchers, and developers must collaborate to create frameworks that prioritize empathy, accessibility, and transparency, ensuring AI systems align with societal values. By integrating interdisciplinary perspectives, stakeholders can design technologies that complement human strengths, rather than replacing or diminishing them. Such efforts are essential not only for fostering trust in AI but also for safeguarding humanity's social and cultural heritage. As technological advancements continue to reshape the world, it is imperative to strike a balance that allows innovation to thrive while preserving the connections, creativity, and traditions that define our shared humanity.

References

[1] I. Abousaber and H. Abdalla, "Review of using technologies of artificial intelligence in companies," *International Journal of Communication Networks and Information Security (IJCNIS)*, 2023. [Online]. Available: https://consensus.app/papers/review-of-using-technologies-of-artificial-intelligence-abousaber-abdalla/6aeed07f7cef5e26b9a6579aa2affa18/?utm_source=chatgpt

[2] V. V. Ramya and A. Khandelwal, "Smart living: The influence of ai on daily activities, day-to-day work and life-style," in *Proceedings of the 2nd ICSSR Conference on "India Towards Viksit Bharat*, vol. 2047, 2024, pp. 13th–14th.

[3] S. Noy and W. Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, vol. 381, pp. 187–192, 2023.

[4] S. Kumar, "Developing human skills in the era of artificial intelligence: Challenges and opportunities for education and training," *Scholedge International Journal of Multidisciplinary Allied Studies*, 2023.

[5] X.-H. Jia and J.-C. Tu, "Towards a new conceptual model of ai-enhanced learning for college students: The roles of artificial intelligence capabilities, general self-efficacy, learning motivation, and critical thinking awareness," *Syst.*, vol. 12, p. 74, 2024.

[6] M. M. Ali, H. M. A. Wafik, S. Mahbub, and J. Das, "Gen z and generative ai: Shaping the future of learning and creativity," *Cognizance Journal of Multidisciplinary Studies*, 2024.

[7] R. Nishant, M. Kennedy, and J. Corbett, "Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda," *International Journal of Information Management*, vol. 53, p. 102104, 2020.

[8] B. W. Susilo and E. Susanto, "Employing artificial intelligence in management information systems to improve business efficiency," *Journal of Management and Informatics*, 2024.

[9] P. Nimkar, D. Kanyal, and S. R. Sabale, "Increasing trends of artificial intelligence with robotic process automation in health care: A narrative review," *Cureus*, vol. 16, 2024.

[10] S. A. Rubab, "Impact of ai on business growth," *The Business and Management Review*, 2023.

[11] V. Bhardwaj, "A systematic review of robotic process automation in business operations: Contemporary trends and insights," *Journal of Intelligent Systems and Control*, 2023.

[12] O. A. Adenekan, N. O. Solomon, P. Simpa, and S. C. Obasi, "Enhancing manufacturing productivity: A review of ai-driven supply chain management optimization and erp systems integration," *International Journal of Management & Entrepreneurship Research*, 2024.

[13] J. Ayers, A. Poliak, M. Dredze, E. Leas, Z. Zhu, J. B. Kelley, D. Faix, A. Goodman, C. Longhurst, M. Hogarth, and D. M. Smith, "Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum," *JAMA Internal Medicine*, 2023.

[14] E. Rich and P. Winston, "Ai in healthcare," *International Journal of Advanced Research in Science, Communication and Technology*, 2024.

[15] Z. Shang, V. Chauhan, K. Devi, and S. Patil, "Artificial intelligence, the digital surgeon: Unravelling its emerging footprint in healthcare – the narrative review," *Journal of Multidisciplinary Healthcare*, vol. 17, pp. 4011–4022, 2024.

[16] S. Mishra, P. Chaudhury, H. K. Tripathy, K. S. Sahoo, N. Jhanjhi, A. A. H. Elnour, and A. Abdelmaboud, "Enhancing health care through medical cognitive virtual agents," *Digital Health*, vol. 10, 2024.

[17] B. Wahn, L. Schmitz, F. N. Gerster, and M. Weiss, "Offloading under cognitive load: Humans are willing to offload parts of an attentionally demanding task to an algorithm," *PLOS ONE*, vol. 18, 2023.

[18] G. Auth, O. Jokisch, and C. Dürk, "Revisiting automated project management in the digital age – a survey of ai approaches," *Online Journal of Applied Knowledge Management*, 2019.

[19] S. A. Hussain, "The integration of artificial intelligence in web accessibility: Enhancing inclusivity," *International Journal for Research in Applied Science and Engineering Technology*, 2024.

[20] I. S. Adeniyi, C. Abimbola, and O. O. Adeleye, "A review of ai-driven pedagogical strategies for equitable access to science education," *Magna Scientia Advanced Research and Reviews*, 2024.

[21] C. Osorio, N. Fuster, W. Chen, Y. Men, and A. Juan, “Enhancing accessibility to analytics courses in higher education through ai, simulation, and e-collaborative tools,” *Inf.*, vol. 15, p. 430, 2024.

[22] K. Chemnad and A. Othman, “Digital accessibility in the era of artificial intelligence—bibliometric analysis and systematic review,” *Frontiers in Artificial Intelligence*, vol. 7, 2024.

[23] A. R. Doshi and O. P. Hauser, “Generative ai enhances individual creativity but reduces the collective diversity of novel content,” *Science Advances*, vol. 10, 2024.

[24] M. Ismayilzada, D. Paul, A. Bosselut, and L. van der Plas, “Creativity in ai: Progresses and challenges,” *ArXiv*, 2024.

[25] R. Patulny and C. Seaman, “‘i’ll just text you’: Is face-to-face social contact declining in a mediated world?” *Journal of Sociology*, vol. 53, pp. 285 – 302, 2017.

[26] E. Venter, “Bridging the communication gap between generation y and the baby boomer generation,” *International Journal of Adolescence and Youth*, vol. 22, pp. 497 – 507, 2017.

[27] M. Chan, “Mobile-mediated multimodal communications, relationship quality and subjective well-being: An analysis of smartphone use from a life course perspective,” *Comput. Hum. Behav.*, vol. 87, pp. 254–262, 2018.

[28] R. Pea, C. Nass, L. Meheula, M. Rance, A. Kumar, H. Bamford, M. Nass, A. Simha, B. Stillerman, S. Yang, and M. Zhou, “Media use, face-to-face communication, media multitasking, and social well-being among 8- to 12-year-old girls,” *Developmental Psychology*, vol. 48, pp. 327–336, 2012.

[29] Z. Mengesha, C. Heldreth, M. Lahav, J. Sublewski, and E. Tuennerman, ““i don’t think these devices are very culturally sensitive.”—impact of automated speech recognition errors on african americans,” *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[30] J. Pezzullo, G. Tung, J. Rogg, L. M. Davis, J. Brody, and W. Mayo-Smith, “Voice recognition dictation: Radiologist as transcriptionist,” *Journal of Digital Imaging*, vol. 21, pp. 384–389, 2008.

[31] A. Lenskjold, J. U. Nybing, C. Trampedach, A. Galsgaard, M. W. Brejnebøl, H. Raaschou, M. Rose, and M. P. Boesen, “Should artificial intelligence have lower acceptable error rates than humans?” *BJR Open*, vol. 5, 2023.

[32] E. Morrow, T. Zidaru, F. Ross, C. Mason, K. Patel, M. Ream, and R. Stockley, “Artificial intelligence technologies and compassion in healthcare: A systematic scoping review,” *Frontiers in Psychology*, vol. 13, 2023.

[33] M. Jeyaraman, S. Balaji, N. Jeyaraman, and S. Yadav, “Unraveling the ethical enigma: Artificial intelligence in healthcare,” *Cureus*, vol. 15, 2023.

[34] A. Fogel and J. Kvedar, “Artificial intelligence powers digital medicine,” *NPJ Digital Medicine*, vol. 1, 2018.

[35] F. Masina, V. Orso, P. Pluchino, G. Dainese, S. Volpato, C. Nelini, D. Mapelli, A. Spagnolli, and L. Gamberini, “Investigating the accessibility of voice assistants with impaired users: Mixed methods study,” *Journal of Medical Internet Research*, vol. 22, 2020.

[36] N. Dolzake, “Review on desktop assistant for visually impaired: Mime.ai,” *International Journal of Scientific Research in Engineering and Management*, 2024.

[37] O. O. Olateju, S. U. Okon, O. O. Olaniyi, A. D. Samuel-Okon, and C. U. Asonze, “Exploring the concept of explainable ai and developing information governance standards for enhancing trust and transparency in handling customer data,” *Journal of Engineering Research and Reports*, 2024.

[38] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, “Trustworthy ai: From principles to practices,” *ACM Computing Surveys*, vol. 55, pp. 1–46, 2021.

[39] C. Siepmann and M. A. Chatti, “Trust and transparency in recommender systems,” *ArXiv*, vol. abs/2304.08094, 2023.

[40] E. O. Sodiya, U. J. Umoga, O. O. Amoo, and A. Atadoga, “Ai-driven warehouse automation: A comprehensive review of systems,” *GSC Advanced Research and Reviews*, 2024.

[41] D. Poleac, “Design thinking with ai,” *Proceedings of the International Conference on Business Excellence*, vol. 18, pp. 2891 – 2900, 2024.

[42] M. H. Massaty, S. K. Fahrurozi, and C. Budiyanto, “The role of ai in fostering computational thinking and self-efficacy in educational settings: A systematic review,” *IJIE (Indonesian Journal of Informatics Education)*, 2024.

[43] R. Merine and S. Purkayastha, “Risks and benefits of ai-generated text summarization for expert level content in graduate health informatics,” *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pp. 567–574, 2022.

[44] J. K. Kato, “The use of ai in enhancing patient monitoring systems,” *Research Output Journal of Public Health and Medicine*, 2024.

[45] M. A. Wahed, M. Alqaraleh, M. Alzboon, and M. S. Al-Batah, “Ai rx: Revolutionizing healthcare through intelligence, innovation, and ethics,” *Seminars in Medical Writing and Education*, 2025.

[46] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education – where are the educators?" *International Journal of Educational Technology in Higher Education*, vol. 16, pp. 1–27, 2019.

[47] E. Elbasi, N. Mostafa, Z. AlArnaout, A. Zreikat, E. Cina, G. Varghese, A. Shdefat, A. Topcu, W. Abdelbaki, S. Mathew, and C. Zaki, "Artificial intelligence technology in the agricultural sector: A systematic literature review," *IEEE Access*, vol. 11, pp. 171–202, 2023.

[48] V. Singh and A. Agnihotri, "Addressing environmental challenges through artificial intelligence (ai)-powered natural disaster management," *International Journal of Applied and Scientific Research*, 2024.

[49] M. Vinichenko, G. Nikiporets-Takigawa, N. V. Ljapunova, O. L. Chulanova, and P. Karácsony, "The nature of the influence of digitalization and artificial intelligence on the sociocultural environment and education in the conditions of the pandemic: views of students of generation z russia and slovakia," *Perspectives of Science and Education*, 2021.

[50] S. Akter, "Global perspectives on the social impacts of artificial intelligence: A comparative review of regional inequalities and cultural contexts," *Journal of Artificial Intelligence General Science (JAIGS)*, 2024.

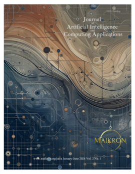
[51] A. Hagerty and I. Rubinov, "Global ai ethics: A review of the social impacts and ethical implications of artificial intelligence," *ArXiv*, vol. abs/1907.07892, 2019.

[52] S. A. A. Kharis and A. Indriyani, "Analyzing social and psychological impacts: Shifting student interaction from teachers to chatgpt in the learning process," *EDUKATIF : JURNAL ILMU PENDIDIKAN*, 2024.

[53] C. Regenbogen, D. A. Schneider, A. Finkelmeyer, N. Kohn, B. Derntl, T. Kellermann, R. Gur, F. Schneider, and U. Habel, "The differential contribution of facial expressions, prosody, and speech content to empathy," *Cognition and Emotion*, vol. 26, pp. 1014–995, 2012.

MAIKRON

Journal
of
Artificial Intelligence
and
Computing Applications



Exploration Paper

Clustering-Based Cyber Situational Awareness: A Practical Approach for Masquerade Attack Detection

Nelva N. Almanza-Ortega[1], **Joaquin Perez-Ortega**[1], **Sergio M. Martinez-Monterrubio**[2], **and Juan A. Recio-Garcia**[3,*]

[1]National Technological Institute of Mexico, México
[2]International University of La Rioja, Spain
[3]University Complutense of Madrid, Spain

ABSTRACT

Cyber Situational Awareness (CSA) is crucial for detecting and mitigating security threats in evolving digital environments. Traditional intrusion detection systems face challenges related to computational efficiency, scalability, and interpretability, particularly in the detection of masquerade attacks, where attackers mimic legitimate user behavior. This exploratory study conducts a preliminary investigation into a clustering-based approach that integrates OK-Means, an optimized variant of K-Means, with k-Nearest Neighbors (k-NN) to improve intrusion detection. The proposed approach is evaluated using the Windows-Users and Intruder Simulations Logs (WUIL) dataset to assess its feasibility and preliminary performance. Experimental results suggest that this method can achieve up to 99% recall in masquerade attack detection while reducing execution time by 85% compared to conventional k-NN classifiers. Additionally, the integration of explainability mechanisms, such as clustering visualization and attack introspection tools, provides security analysts with interpretable insights into system decisions. As an initial exploration, this study provides early-stage insights into clustering-based CSA methods and lays the groundwork for future research. The findings suggest that this approach can be further developed and extended to other cybersecurity domains, such as phishing and malware detection, contributing to AI-driven security frameworks.

Keywords: cyber situational awareness (CSA), masquerade attack detection, explainable machine learning

1. Introduction

Cyber Situational Awareness (CSA) plays a crucial role in protecting IT systems against evolving cyber threats. As organizations increasingly rely on digital infrastructure, the need for effective intrusion detection mechanisms has become more critical than ever [1]. Traditional intrusion detection systems (IDSs) often struggle with two significant challenges: the complexity of big data environments and the lack of interpretability in decision-making processes [2]. These challenges are especially pronounced in the detection of masquerade attacks, where malicious actors disguise their activity as that of legitimate users. The dynamic and high-volume nature of cybersecurity logs further complicates the real-time identification of such threats, making conventional predictive models less effective in adaptive environments.

To address these limitations, this work explores a clustering-based approach that enhances CSA by

E-mail address: jareciog@fdi.ucm.es
https://doi.org/10.5281/zenodo.14933955

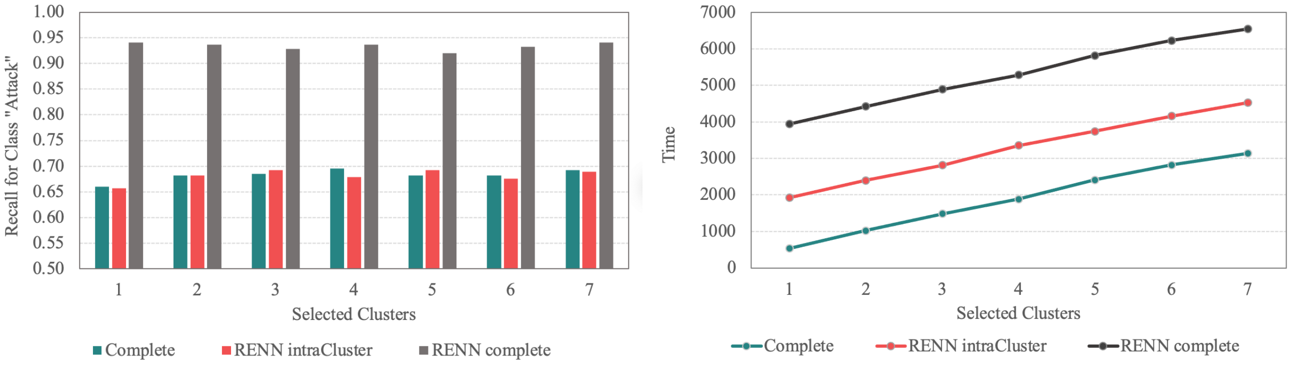


Figure 1. Recall (left) and processing time (right) using different noise reduction approaches (dataset 20%).

improving both efficiency and explainability in masquerade attack detection. Our method combines OK-Means, an optimized variant of K-Means [3], with k-Nearest Neighbors (k-NN) to streamline the classification of potential intrusions. This hybrid approach reduces computational cost while maintaining high detection performance, making it more suitable for real-time threat analysis. Additionally, we integrate explainability strategies to provide security analysts with transparent, interpretable insights into why a particular behavior is flagged as suspicious, improving decision-making in CSA.

This study presents an experimental evaluation of the proposed method using the Windows-Users and Intruder Simulations Logs dataset (WUIL) [4]. We analyze the feasibility and early-stage performance of clustering-based detection in reducing false negatives while maintaining high accuracy. As an exploratory study, this paper provides practical insights into the implementation, optimization, and real-world applicability of AI-driven cybersecurity solutions. These findings serve as a foundation for future research into adaptive and scalable intrusion detection models.

2. Project Description

The proposed approach enhances CSA by improving the efficiency and interpretability of masquerade attack detection. The method leverages a combination of OK-Means clustering [3] and k-Nearest Neighbors (k-NN) to classify potential intrusions while optimizing computational resources. Clustering reduces the search space by grouping similar user behaviors, allowing k-NN to perform instance-based classification on a smaller, more relevant subset of data. This structure improves detection efficiency without significantly compromising accuracy.

OK-Means is a refinement of the traditional K-Means algorithm that optimizes cluster updates, making it better suited for dynamic cyber environments where system behavior constantly evolves. Unlike static machine learning models that require frequent retraining, OK-Means efficiently adapts to new patterns while maintaining clustering quality. Furthermore, k-NN provides an explainable classification process, enabling security analysts to understand why an alert was triggered. To further enhance accuracy, the system integrates Repeated Edited Nearest Neighbor (RENN) [5], a noise reduction technique that filters out inconsistencies and redundant data, reducing false positives and improving model reliability.

The implementation consists of three key steps. First, user activity data is collected from system logs using User Activity Monitoring (UAM) sensors [4]. This data is then preprocessed to extract spatial, temporal, and directional locality features, which help characterize user behavior. Second, the clustering model is periodically updated based on data volume and classification error rate to adapt to evolving attack patterns. Finally, explainability is enhanced through visual analysis tools that allow security analysts to inspect cluster structures and attack feature distributions, making CSA more interpretable and actionable.

3. Implementation and Results

The effectiveness of the proposed clustering-based approach was evaluated using real-world data, with a focus on improving detection efficiency and explainability. This section details the dataset, experimental setup, performance improvements, and the visual tools designed to aid security analysts.

3.1 Dataset and Preprocessing

The evaluation was conducted using the Windows-Users and Intruder Simulations Logs dataset (WUIL) [4], which contains a total of 54,649 instances. The dataset includes both legitimate user activities and masquerade attack attempts, making it well-suited for intrusion detection research. Due to the nature of masquerade attacks, the dataset is highly imbalanced, with 96.77% of instances belonging to legitimate users and only 3.33% corresponding to attacks.

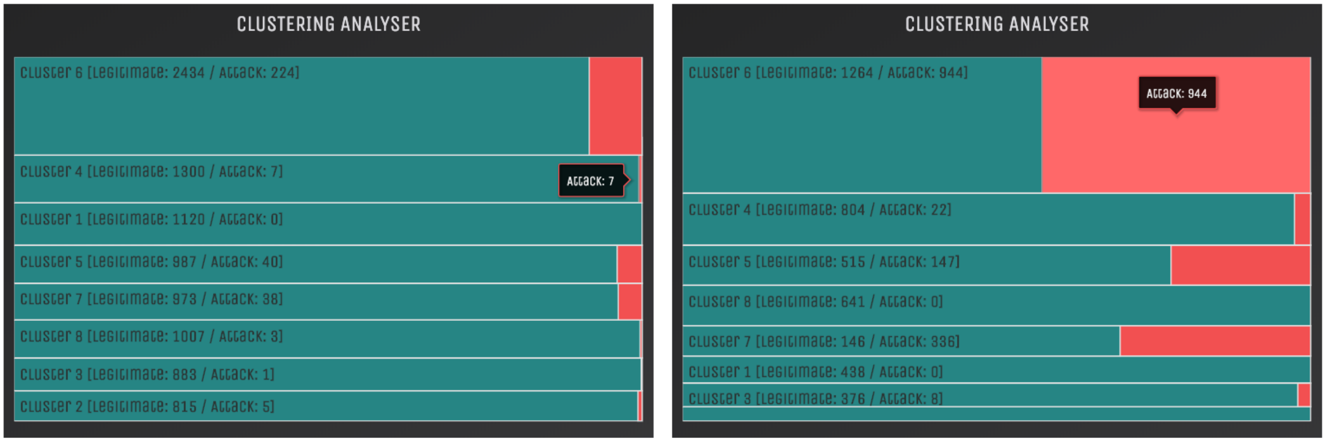

CLUSTERING ANALYSER
Cluster 6 [Legitimate: 2434 / Attack: 224]
Cluster 4 [Legitimate: 1300 / Attack: 7]
Attack: 7
Cluster 1 [Legitimate: 1120 / Attack: 0]
Cluster 5 [Legitimate: 987 / Attack: 40]
Cluster 7 [Legitimate: 973 / Attack: 38]
Cluster 8 [Legitimate: 1007 / Attack: 3]
Cluster 3 [Legitimate: 883 / Attack: 1]
Cluster 2 [Legitimate: 815 / Attack: 5]
CLUSTERING ANALYSER
Cluster 6 [Legitimate: 1264 / Attack: 944]
Attack: 944
Cluster 4 [Legitimate: 804 / Attack: 22]
Cluster 5 [Legitimate: 515 / Attack: 147]
Cluster 8 [Legitimate: 641 / Attack: 0]
Cluster 7 [Legitimate: 146 / Attack: 336]
Cluster 1 [Legitimate: 438 / Attack: 0]
Cluster 3 [Legitimate: 376 / Attack: 8]

Figure 2. Clustering analyzer tool when visualizing the 20% subsampling (left) and the complete dataset (right).

The dataset consists of multiple behavioral features extracted from user interactions with the file system. These features include:

- File access frequency.
- Event distances between accessed files.
- Directionality of file system navigation.
- Temporal locality patterns in access sequences.

These attributes provide a structured representation of user behavior, enabling clustering and classification models to distinguish between normal and anomalous activities.

3.2 Experimental Setup and Configuration

To analyze the performance of the proposed method, we conducted multiple experiments with different configurations of the clustering and classification components:

- Clustering was performed with $C = 4$ and $C = 8$ clusters.
- For classification, we tested k-NN with $k = 1, 3, 5$ neighbors.
- The number of selected clusters for classification (sC) was varied as $sC = 1, 2, 4$.
- The impact of noise reduction was evaluated using Repeated Edited Nearest Neighbor (RENN) [5], which filters out inconsistent or redundant samples.

The evaluation process included 10-fold cross-validation on five subsamples of the dataset, using 20%, 40%, 60%, 80%, and 100% of the instances. These configurations allowed us to assess the balance between computational efficiency and detection accuracy.

3.3 Performance Gains

The experimental results demonstrated substantial improvements in both computational efficiency and detection performance. Figure 1 presents the recall values and execution times for different noise reduction and clustering strategies.

- **Time Efficiency:** By leveraging OK-Means, execution time was reduced by up to 85% compared to a standard k-NN classifier without clustering.
- **Detection Performance:** The combination of OK-Means and RENN significantly enhanced classification accuracy, achieving up to 99% recall for the detection of masquerade attacks.
- **False Negative Reduction:** The clustering approach minimized the number of undetected attacks, improving the overall reliability of the system.

3.4 Explainability and Visual Tools

A key advantage of this approach is the integration of explainability mechanisms, allowing security analysts to interpret the reasoning behind attack classifications. Two main visual tools were developed:

- **Clustering Analyzer:** This tool provides a hierarchical tree-map visualization of the dataset, showing the proportion of legitimate and attack instances in each cluster. The tool helps analysts assess the risk level of detected attacks by highlighting clusters where anomalies are found.
- **Attack Introspection Tool:** This visualization tool presents a polar chart comparing the detected attack's feature values with those of similar past attacks and the cluster prototype. This allows analysts to understand which behavioral attributes contributed to the classification.

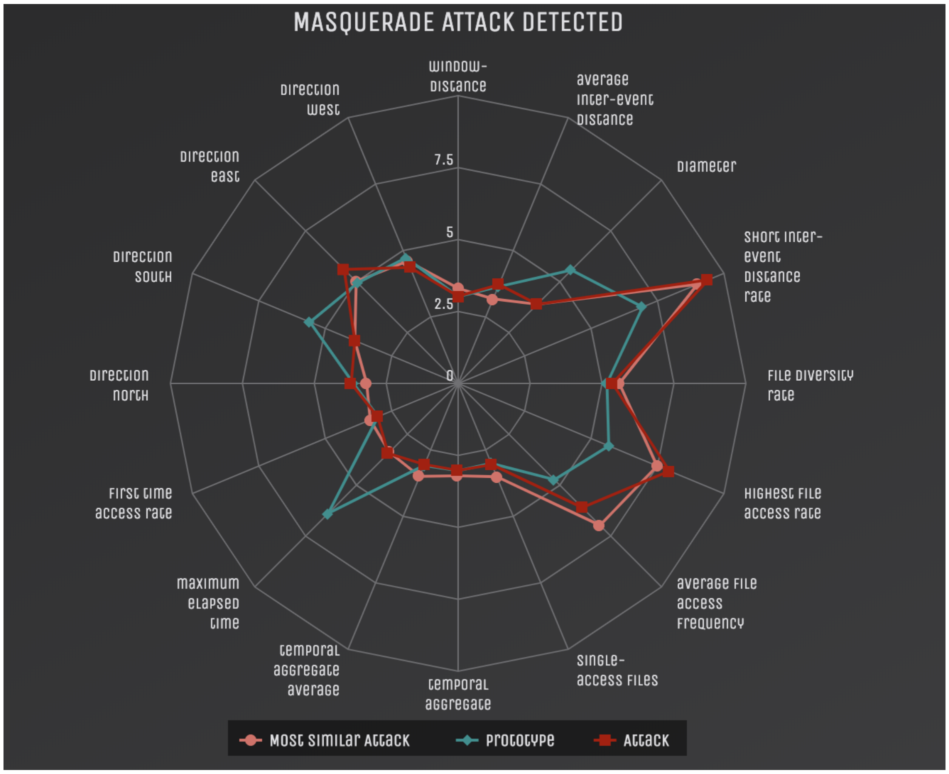


Figure 3. Screenshot of the attack introspection tool showing the features of the attack, the most similar attack that raised the alarm, and the cluster's prototype.

These tools facilitate CSA by enabling analysts to verify whether detected threats are legitimate or false positives, improving trust in automated intrusion detection systems.

4. Discussion and Potential Impact

The proposed clustering-based approach for masquerade attack detection provides a balance between accuracy, efficiency, and interpretability, which are critical components of real-world CSA. Unlike traditional black-box machine learning models, this method offers security analysts actionable insights by leveraging an explainable classification process. The integration of OK-Means and k-NN enhances real-time threat detection while reducing computational overhead, making it suitable for large-scale cybersecurity applications.

While the initial findings demonstrate promising improvements in efficiency and detection performance, further validation is needed to assess long-term scalability and real-world deployment challenges. Future research directions involve integrating adaptive learning techniques to dynamically update clusters based on evolving attack patterns. This would further improve the model's ability to detect previously unseen threats. Additionally, the approach can be expanded to other cybersecurity domains, such as phishing detection, malware analysis, and anomaly detection in industrial control systems. By refining the clustering process and incorporating hybrid AI techniques, this method has the potential to contribute to broader cybersecurity frameworks that prioritize both performance and interpretability.

5. Conclusion

This study explores the feasibility of a clustering-based approach for CSA in detecting masquerade attacks. By combining OK-Means with k-NN, the approach demonstrates a potential balance between computational efficiency, detection accuracy, and explainability. Preliminary results suggest that this method can detect up to 99% of masquerade attacks while reducing execution time by up to 85%. Furthermore, the integration of visual tools enhances interpretability, allowing security analysts to make informed decisions based on attack classifications.

As an exploratory investigation, this work provides early-stage insights into the advantages and limitations of clustering-based CSA methods. Future research should focus on further validation in large-scale, real-world environments, as well as integrating adaptive learning techniques to improve robustness against evolving threats. These findings lay the groundwork for continued innovation in AI-driven cybersecurity solutions.

References

[1] M. Endsley, “Endsley, m.r.: Toward a theory of situation awareness in dynamic systems. human factors journal 37(1), 32-64,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 37, pp. 32–64, 03 1995.

[2] M. Conti, A. Dehghantanha, and T. Dargahi, “Cyber threat intelligence : Challenges and opportunities,” *CoRR*, vol. abs/1808.01162, 2018. [Online]. Available: http://arxiv.org/abs/1808.01162

[3] J. Pérez-Ortega, N. N. Almanza-Ortega, and D. Romero, “Balancing effort and benefit of k-means clustering algorithms in big data realms,” *PloS one*, vol. 13, no. 9, p. e0201874, 2018.

[4] J. B. Camiña, C. Hernandez-Gracidas, R. Monroy, and L. Trejo, “The windows-users and -intruder simulations logs dataset (wuil): An experimental framework for masquerade detection mechanisms,” *Expert Systems with Applications*, vol. 41, no. 3, pp. 919 – 930, 2014, methods and Applications of Artificial and Computational Intelligence. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417413006349

[5] D. Hand and B. Batchelor, “Experiments on the edited condensed nearest neighbor rule,” *Information Sciences*, vol. 14, no. 3, pp. 171 – 180, 1978. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0020025578900403