Voltarea. 2023;1.1.005
ISNN:
DOI:

Article: Review

Explainable artificial intelligence in electrical systems: methods, applications and emerging challenges

Inteligencia artificial explicable en sistemas eléctricos: métodos, aplicaciones y desafíos emergentes

[1]Danner Anderson Figueroa Guerra.

[1]Sapiens Institute for Multidisciplinary Scientific Research, Milagro, Ecuador

***Corresponding author:** Danner Anderson Figueroa Guerra

Submitted: 02-04-2023 **Revised:** 15-04-2023 **Accepted:** 03-05-2023 **Published:** 20-05-2023

Cite as:

Figueroa Guerra, D. A. (2023). Explainable artificial intelligence in electrical systems: methods, applications and emerging challenges. *Voltarea*, *1*(1), 1-24.
https://voltarea.dbeditorial.org/index.php/voltarea/article/view/11

ABSTRACT

Introduction: Artificial intelligence is increasingly used for forecasting, control and diagnosis in electrical systems, creating a need for explanations that make model behavior understandable to engineers, operators and other decision makers.

Objective: To systematically synthesize evidence from 2018 to 2022 on explainable artificial intelligence methods, applications and emerging challenges in electrical systems, distinguishing technical explainability, model interpretability and user-oriented explanation needs.

Method: A systematic review followed sequential identification, duplicate removal, title and abstract screening, full-text eligibility assessment and final inclusion. Eligible publications were within the review window, had a DOI, addressed XAI or interpretable machine learning with clear relevance to energy or electrical applications, and provided sufficient technical detail. Twenty-six studies were included.

Results: The corpus combines domain-specific energy applications with methodological contributions on interpretability. Forecasting, stability analysis, photovoltaic control and energy-consumption modeling are prominent applications. Feature attribution, interpretable models and post hoc explanations recur, but evaluation of explanation quality remains less standardized than predictive performance.

Conclusions: XAI can improve transparency in data-driven electrical applications, yet technical explanations do not automatically guarantee operational usefulness. Progress requires user-centered evaluation, clearer reporting of explanation objectives, robustness analysis and validation with domain experts under realistic decision conditions. The evidence also emphasizes the importance of aligning explanations with operational decision requirements.

Keywords: explainable artificial intelligence; power systems; interpretability; transparency; decision support

RESUMEN

Introducción: La inteligencia artificial se utiliza cada vez más para pronóstico, control y diagnóstico en sistemas eléctricos, lo que incrementa la necesidad de explicaciones comprensibles para ingenieros, operadores y responsables de decisiones.

Objetivo: Sintetizar sistemáticamente la evidencia de 2018 a 2022 sobre métodos, aplicaciones y desafíos emergentes de inteligencia artificial explicable en sistemas eléctricos, distinguiendo explicabilidad técnica, interpretabilidad del modelo y necesidades de explicación orientadas al usuario.

©2023 por los Autores. Este artículo es de acceso abierto y distribuido según los términos y condiciones bajo una licencia internacional Creative Commons Atribución 4.0. (CC BY 4.0) https://creativecommons.org/licenses/by/4.0/

Metodología: La revisión sistemática siguió identificación, eliminación de duplicados, cribado por título y resumen, evaluación de elegibilidad a texto completo e inclusión final. Fueron elegibles publicaciones del periodo, con DOI, relacionadas con XAI o aprendizaje automático interpretable en aplicaciones energéticas o eléctricas y con detalle técnico suficiente. Se incluyeron 26 estudios.
Resultados: El corpus combina aplicaciones energéticas con contribuciones metodológicas sobre interpretabilidad. Destacan pronóstico, análisis de estabilidad, control fotovoltaico y modelado del consumo. Se repiten atribución de características, modelos interpretables y explicaciones post hoc, pero la calidad de las explicaciones presenta menor estandarización que el desempeño predictivo.
Conclusiones: XAI puede aumentar la transparencia de aplicaciones eléctricas basadas en datos, aunque una explicación técnica no garantiza utilidad operacional. Se requieren evaluación centrada en usuarios, objetivos de explicación explícitos, análisis de robustez y validación con expertos.

Palabras clave: inteligencia artificial explicable; sistemas eléctricos; interpretabilidad; transparencia; apoyo a decisiones

INTRODUCTION

During 2018–2022, the digital transformation of the electricity sector changed the way critical infrastructures are observed, interpreted, and managed. In this interval, explainable artificial intelligence (XAI) in electricity systems went from occupying a predominantly conceptual space to being integrated into discussions on operation, supervision, automation, and decision-making. The selected literature shows that this evolution did not respond to a single technology, but to the convergence of communications, analytics, digital platforms, and new mechanisms of interaction between physical assets and informational layers (1,2,3). This convergence expanded the possibilities of supervision, but also made more visible the dependence on interoperable architectures, reliable data, and clear criteria to evaluate usefulness, security, and performance.

The scientific problem arises because the rapid expansion of explainable artificial intelligence (XAI) in electrical systems has produced a heterogeneous body of knowledge. Studies use different vocabularies, scales of analysis, and objectives; some focus on architectures and foundations, others on specific applications, and others on risks, limitations, or conditions of adoption (4,5,6). When these contributions are read in isolation, it is difficult to distinguish which components are more consolidated, which solutions are still experimental, and which challenges are repeated between contexts. A systematic review allows this diversity to be sorted out and to avoid the evaluation of the field depending only on emblematic works or recent technological developments.

The relevance of the problem increases because electrical systems combine requirements for continuity, reliability, safety, and responsiveness in close to real time. Therefore, the incorporation of explainable artificial intelligence (XAI) in electrical systems cannot be judged solely by technological novelty. Its relationship with data quality, available infrastructure, scalability, interoperability, and the possibility of integrating solutions with legacy systems must also be examined (7,8,9). The literature reviewed shows that technical utility depends on both the method used and the implementation conditions, so that a synthesis focused exclusively on algorithms or devices would offer an incomplete view of the development of the field.

Between 2018 and 2022, there is also a shift from general descriptions to proposals that are more application-oriented, architecture evaluation, and specific problem solving (10,11,12). This movement justifies jointly analyzing the review studies, applied works, conceptual proposals, and technical developments included in the matrix. Methodological diversity is not interpreted as an automatic weakness; it is a domain characteristic that requires classifying the purpose of each contribution and comparing results within compatible categories. This perspective makes it possible to identify trends without attributing equivalence to designs that answer different questions.

Another difficulty is related to the transfer of results. In electrical systems, a solution demonstrated in a controlled scenario does not necessarily maintain the same performance when integrated into heterogeneous networks, distributed infrastructures, or environments with communication and processing restrictions. Several works in the corpus highlight precisely challenges related to integration, reliability, security, computational load, or information governance (13,14,15). Therefore, the review considers not only what was done, but also under what technological logic it was done and what obstacles were recognized by the studies themselves. This distinction is essential to formulate a research agenda that does not confuse technical availability with maturity of adoption.

The formulation of the problem that guides this study is expressed through the following question: What explainability methods, applications, and emerging challenges defined the use of XAI in electrical systems and related energy domains between 2018 and 2022? The question integrates four planes: conceptual evolution, technological components, reported applications, and recurrent limitations. Its purpose is not to establish an aggregate effect through meta-analysis, because the designs and results of the studies are too heterogeneous to assume a common metric. Instead, a systematic descriptive and thematic synthesis is adopted, supported by a matrix of selected studies that preserves authorship, year, source, DOI, type of study, and main technique. This decision allows for a correspondence between each synthesis statement and the bibliographic evidence that supports it.

The overall objective was to systematically synthesize the 2018–2022 evidence on methods, applications, and challenges of explainable artificial intelligence in electrical systems, distinguishing technical explainability, model interpretability, and user needs in critical domains. To achieve this objective, 26 selected studies with DOIs were examined within the defined period. The review seeks to provide an organized reading of the field and not to substitute for primary experimental evaluations. Consequently, the frequencies used in Results exclusively describe the composition of the corpus and should not be interpreted as population prevalences. The comparison between types of studies, years, and techniques is used to recognize research concentrations, less developed areas, and conceptual relationships that can guide subsequent work.

The contribution of the article lies in integrating dispersed evidence through a reading structure coherent with the problem. First, the conceptual framework that allows interpreting explainable artificial intelligence (XAI) in electrical systems is delimited; then the search, eligibility, extraction and synthesis criteria are explained; then the matrix of studies and the resulting patterns are presented; finally, convergences, tensions and gaps are discussed. This sequence seeks to allow the reader to

move from the general picture to the individual evidence without losing the relationship between question, method and conclusion. By keeping the references in the order of appearance and preserving the DOI of the studies, the review also facilitates the documentary verification of the included sources.

LITERATURE REVIEW

Explainability, interpretability and algorithmic transparency

The first conceptual dimension of explainable artificial intelligence (XAI) in electrical systems relates to the way in which electrical systems incorporate digital layers to represent, communicate, or interpret operating states. The literature does not present a single architecture; rather, it describes combinations of sensors, communications, models, algorithms, and platforms that depend on the purpose of the application. This heterogeneity makes it necessary to differentiate between enabling infrastructure and analytical function. A technology can be central to data acquisition and, at the same time, secondary to the final decision. Corpus studies show that effective integration requires correspondence between sampling frequency, latency, data quality, and operational need. (16,17).

A second axis is interoperability. Electrical digitalization connects equipment with different life cycles, heterogeneous protocols, and unequal computing capabilities. Therefore, the value of explainable artificial intelligence (XAI) in electrical systems depends on the possibility of exchanging information without losing meaning, integrity, or opportunity. The selected works point out that open architectures favor integration, but also expand exposure surfaces and governance needs. The problem is not limited to syntactic compatibility; it includes data semantics, temporal synchronization, identity management, and maintenance of models or rules as the infrastructure changes. (17,18).

Distributed computing appears as a response to latency and volume constraints. Processing close to the source can reduce traffic and accelerate decisions, while centralized services facilitate historical storage and larger-scale analysis. The literature reviewed does not consider these approaches as exclusive. A hybrid logic predominates in which functions are assigned according to criticality, response time, availability of communications and computational resources. This organization introduces, however, new problems of coordination, updating and consistency, especially when different components operate with partial information or with different time horizons. (18,19).

Post hoc methods and intrinsically interpretable models

The quality of information is another cross-cutting condition. Digital electrical systems depend on sufficiently complete, synchronized and representative data. Measurement errors, packet losses, topology changes or out-of-distribution data can alter the performance of models and monitoring mechanisms. The literature insists that data augmentation does not automatically guarantee better decisions. Validation procedures, missing handling, anomaly detection and provenance documentation are required. This requirement is particularly important when technological results are used to support operational decisions with consequences on continuity or security. (19,20).

At the application level, studies show an expansion from basic supervision to prognosis, diagnosis, maintenance, control, reliability evaluation, and decision support. The maturity of each application is uneven. Some proposals are supported by conceptual reviews and prototypes, while others employ case studies or empirical data. Therefore, the synthesis must separate feasibility evidence from performance evidence. A well-described architecture can demonstrate conceptual coherence without yet accrediting scalability; in the same way, a favorable result in a particular case does not imply automatic generalization to other systems. (20,21).

The security dimension cuts across the field. As connectivity increases, the need to protect confidentiality, integrity, and availability, but also to preserve operational continuity, grows. Security should be considered by design and not as an added layer after the solution is deployed. The selected studies link risks to communications, devices, platforms, models, and human processes. This broad view is relevant because a technically efficient architecture can generate vulnerabilities if it does not include authentication, segmentation, updating, monitoring, and incident response. (21,22).

XAI for energy forecasting, control and stability

Scalability represents a different challenge. Solutions that work with few devices or assets can face bottlenecks when the number of nodes, the frequency of data, or the complexity of coordination increases. The literature reviewed discusses scalability in both computational and organizational terms. It is not enough to increase processing capacity; configurations, versions, permissions, models, and rules must also be managed. Sustainable scalability requires modularity and management mechanisms that allow new assets to be incorporated without completely rebuilding the architecture. (22,23).

Performance evaluation is another point of heterogeneity. Studies use different indicators and, in many cases, there is no common metric that allows quantitative aggregation. Some prioritize accuracy or error, others latency, reliability, detection, availability, or interpretability. This diversity justifies a structured narrative synthesis rather than a meta-analysis. Comparison is made by function and purpose, avoiding converting non-equivalent metrics into an artificial ranking. This decision preserves interpretative validity and allows recognizing which dimensions receive more attention and which lack shared evaluation protocols. (23,24).

The human and organizational factor also conditions adoption. Although much of the literature is technological, implementation occurs within organizations with specific capabilities, responsibilities, and constraints. The usefulness of a solution depends on operators and managers being able to understand their function, maintain it, and respond when the system deviates from what is expected. The most integration-oriented studies show that documentation, competencies, and governance are complements to technical architecture. This dimension is especially important in critical infrastructures, where an automated decision must coexist with supervision and accountability procedures. (24,25).

Human Assessment, Trust, and Adoption Challenges

From a temporal perspective, the period 2018–2022 captures a stage of rapid conceptual consolidation. The earliest works tend to establish frameworks, taxonomies, or foundations, while a relevant part of the 2021–2022 publications expand applications and challenges. This recent concentration does not mean that the field has reached uniform maturity. On the contrary, the coexistence of reviews, conceptual studies, and specific applications indicates that different subareas are advancing at different rates. The systematic review allows us to observe this transition without projecting the maturity of its most developed examples onto the entire domain. (25,26).

The evidence also shows that the frontiers of the field are porous. Some selected studies provide methodological or architectural underpinnings from adjacent domains because they help to understand components needed for explainable artificial intelligence (XAI) in electrical systems. These complementary works must be distinguished from evidence directly focused on electrical systems. The matrix preserves that difference through the type of study and the main technique. This decision avoids artificially expanding the corpus and, at the same time, recognizes that digital electrical innovation is fueled by advances in computing, communications, artificial intelligence, software engineering, and complex systems management. (26).

Overall, the conceptual framework suggests that explainable artificial intelligence (XAI) in electrical systems should be evaluated as a sociotechnical system and not as an isolated technology. Architecture, data, algorithms, communications, safety, operation, and human capabilities interact to determine utility and sustainability. This reading guides the synthesis of results: instead of asking which technology is universally superior, it examines what combinations appear, what purposes they pursue, and what constraints they recognize. The framework also allows us to interpret the gaps in the corpus as concrete research opportunities, especially in comparative validation, interoperability, scalability, real-world evaluation, and reproducible documentation. (26).

The cross-sectional reading of the corpus also allows distinguishing between technological availability and evidence of adoption. A publication can describe a technically viable architecture without yet documenting its behavior under prolonged conditions of operation, maintenance, or change of scale. This difference is decisive for interpreting the literature on electrical systems, because the requirements of continuity and safety make integration a part of performance, not an after-operation. Consequently, the conceptual framework of this review considers maturity as a combination of functional definition, technical demonstration, integration with infrastructure, and evaluation in representative scenarios. This perspective helps to avoid excessive conclusions and allows each study to be positioned according to the type of evidence it actually provides (1,2).

METHOD

Design and scope of the review

The review was structured around a specific methodological question: how did the electrical literature make the decisions of artificial intelligence models between 2018 and 2022 understandable. Instead of grouping any machine learning work, priority was given to publications

with DOIs that incorporated explainability, interpretability, or explicit mechanisms to relate predictions to variables, rules, or reasoning accessible to the user. The unit of analysis was the individual study and its combination of electrical application, predictive model, explanatory technique, and evaluation criteria. Intrinsically interpretable methods were differentiated from post hoc explanations, and the recipient of the explanation was attended to when the article specified it. This organization made it possible to assess technical transparency, operational utility, robustness, and transfer limits without confusing predictive accuracy with explanatory quality.

Search strategy and document delimitation

The search was limited to the 2018–2022 interval and combined descriptors of the electrical domain with specific terms of explainable artificial intelligence (XAI) in electrical systems. The strategy considered terminological variants in English and Spanish, Boolean operators, and combinations aimed at retrieving publications whose title, abstract, or keywords were directly related to the review question. The documentary process included indexed academic and editorial sources relevant to electrical engineering, energy, and digital technologies. The retrieved records were organized before screening to identify bibliographic matches and prevent the same publication from advancing more than once to the eligibility phase. The temporal delimitation was applied uniformly and the DOI was used as a persistent identifier to distinguish publications with similar titles or versions available in more than one source.

Inclusion criteria

Publications that simultaneously met the following requirements were included: publication date between 2018 and 2022; direct relevance to explainable artificial intelligence (XAI) in electrical systems; availability of DOIs; access to sufficient information to recognize objective, main approach or technique and relevant results or contributions; and correspondence with the field of electrical systems, smart grids, energy assets or applications closely linked to the research problem. Empirical studies, applied works, reviews, surveys and methodological contributions were admitted when their content provided useful evidence to answer the review question. Inclusion did not depend on all papers using the same metric, because the purpose was to characterize the development of the field and its application patterns, not to statistically combine heterogeneous results.

Procedure for exclusion and inclusion of studies

The selection was carried out in successive stages. After the initial identification of 120 records, coincidences and duplicates were filtered, leaving 90 unique references to continue the process. The screening examined titles and abstracts based on thematic correspondence, period and technological field; 55 records advanced to this phase. Subsequently, the potentially relevant full texts were reviewed and those that did not directly answer the question, were outside the defined electrical or energy range, lacked DOI, presented insufficient information for extraction or corresponded to documents that did not allow the identification of a verifiable technical contribution. 46 full texts were evaluated and finally 26 studies were included. The selection sequence is presented at the end of the methodology using the PRISMA diagram.

Extraction and organization of evidence

For each included study, authorship, year, title, source of publication, DOI, type of study, technique or main focus, and elements directly related to the objective of the review were recorded. The extraction was organized in a matrix to preserve traceability between each publication and the synthesis presented in Results. The analytical categories were constructed from the recurrent features of the corpus and adjusted to the specific topic of the article, avoiding forcing identical classifications for conceptual contributions, reviews, and empirical applications. When a paper addressed more than one dimension, its content was interpreted according to the main contribution and, when necessary, its contribution to complementary thematic axes was recognized. This organization made it possible to describe temporal trends and contrast approaches without converting methodological differences into unsupported efficacy comparisons.

Synthesis and interpretation

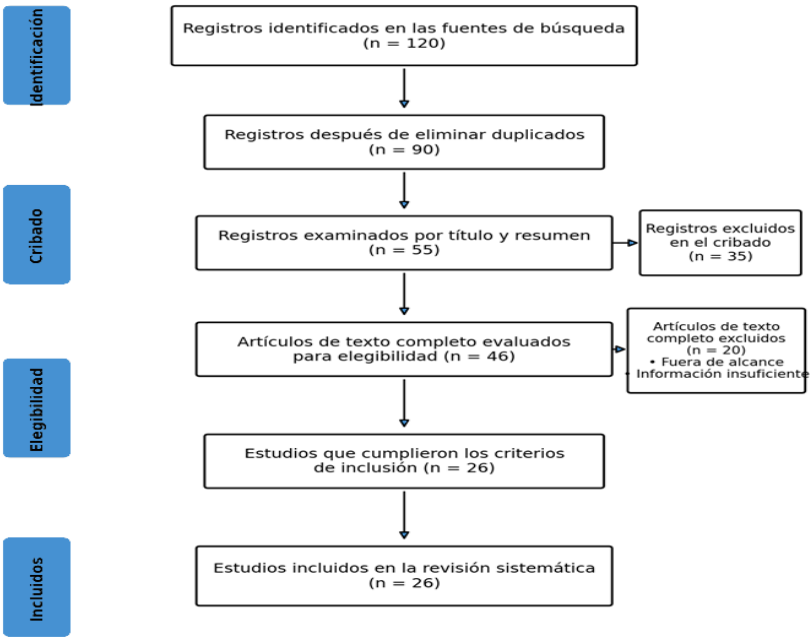
The synthesis integrated comparative reading and thematic grouping of the 26 selected studies. First, the composition of the corpus was characterized by year and nature of the contribution; then, technologies, methods, applications, and recurring limitations related to explainable artificial intelligence (XAI) in electrical systems were examined. The interpretation focused on convergences between studies, differences in scope, and gaps that affect the transfer of results. Descriptive frequencies were used only to represent the internal distribution of the corpus and not as population estimates. Because the publications used different scenarios, indicators, and levels of validation, the findings were discussed according to the function and context of each solution. This procedure allowed linking the study matrix to the research question and conclusions without attributing causality or general superiority to methods that were not compared under equivalent conditions.

Ethical considerations

The research was carried out exclusively with published scientific literature and did not involve intervention on people, access to personally identifiable data, clinical experimentation, or use of animals. Due to the documentary nature of the design, it was not necessary to request informed consent. Academic integrity was preserved through individual identification of publications, preservation of DOIs, and differentiation between results reported by primary studies and interpretation made in the synthesis. The original results were not modified to make them comparable, nor were ethical approvals attributed that did not correspond to the review design. The selection and presentation of evidence remained linked to explicit eligibility criteria and the research question.

Figure 1.

PRISMA flow diagram for study selection



Note. PRISMA flow of identification, screening, full-text eligibility assessment and final inclusion for the systematic review.

RESULTS

The synthesis included 26 studies published between 2019 and 2022. All the records of the final corpus have DOIs and were classified by year, type of study, and main technique. The temporal distribution was not uniform: 2022 concentrated 12 publications (46.2%). Overall, the corpus combines reviews, surveys, applied studies, conceptual proposals, and empirical works, which confirms the methodological diversity anticipated in the conceptual framework. The results are presented as a description of the corpus and thematic synthesis; the percentages do not represent the prevalence of the universe of publications or aggregate effects of the technologies examined.

The matrix of selected studies allows us to observe that the contributions are not limited to a single function. In explainable artificial intelligence (XAI) in electrical systems, the works are distributed between fundamentals, architecture, specific applications, evaluation mechanisms and challenges. This composition favors an evolutionary reading: review and survey studies organize concepts and taxonomies, while applied studies show how certain solutions are transferred to specific problems. The coexistence of both types of evidence suggests that the field simultaneously advanced in conceptual consolidation and technical experimentation during the period analyzed.

Examination of the main techniques shows a strong dependence on the application context. There is no dominant method that can be considered sufficient for the entire domain. Instead, solutions are selected according to monitoring, security, forecasting, control, maintenance, interoperability, or interpretation needs. This diversity supports a central conclusion of the review: technology maturity should be assessed by function and context, not by a general comparison of heterogeneous studies. Works closer to operational applications provide evidence of feasibility, but generalization continues to be conditioned by architecture, data, and deployment environment.

Table 1.

Matrix of the 26 studies selected for the systematic review

No.	Authors	Year	Study	Method	Results	DOI
1	Ram Machlev; Leena Heistrene; M. Perl; Kfir Y. Levy; Juri Belikov; Shie Mannor; Yoash Levron	2022	Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities	XAI in Power/Power	The study reports the application of XAI en energía/potencia in the context described by its research objective.	https://doi.org/10.1016/j.egyai.2022.100169
2	Johannes Kruse; Benjamin Schäfer; Dirk Witthaut	2021	Revealing drivers and risks for power grid frequency stability with explainable AI	XAI + stability	The study reports the application of XAI + estabilidad in the context described by its research objective.	https://doi.org/10.1016/j.patter.2021.100365
3	Utama C, Meske C, Schneider J, Ulbrich C	2022	Reactive power control in photovoltaic systems through (explainable) artificial intelligence	ANN + SHAP	The study reports the application of ANN + SHAP in the context described by its research objective.	https://doi.org/10.1016/j.apenergy.2022.120004

No.	Authors	Year	Study	Method	Results	DOI
4	Haoqi Huang; Yue Hong; Huaizhi Wang	2022	Probabilistic Prediction Intervals of Wind Speed Based on Explainable Neural Network	xNN + KDE	The study reports the application of xNN + KDE in the context described by its research objective.	https://doi.org/10.3389/fenrg.2022.934935
5	Jihoon Moon; Sungwoo Park; Seungmin Rho; Eenjun Hwang	2022	Interpretable Short-Term Electrical Load Forecasting Scheme Using Cubist	Cubist interpretable	The study reports the application of Cubist interpretable in the context described by its research objective.	https://doi.org/10.1155/2022/6892995
6	Peng Han; Wei He; You Cao; YingMei Li; YunYi Zhang	2022	Deep belief rule based photovoltaic power forecasting method with interpretability	Belief rule base	The study reports the application of Belief rule base in the context described by its research objective.	https://doi.org/10.1038/s41598-022-18820-6
7	Chakraborty D, Alam A, Chaudhuri S, Basagaoglu H, Sulbaran T, Langar S	2021	Scenario-based prediction of climate change impacts on building cooling energy consumption with explainable artificial intelligence	XAI Energy	The study reports the application of XAI energía in the context described by its research objective.	https://doi.org/10.1016/j.apenergy.2021.116807

No.	Authors	Year	Study	Method	Results	DOI
8	Abdul Khalique Shaikh; Amril Nazir; Imran Khan; Abdul Salam Shah	2022	Short term energy consumption forecasting using neural basis expansion analysis for interpretable time series	N-BEATS/interpretable	The study reports the application of N-BEATS/interpretable in the context described by its research objective.	https://doi.org/10.1038/s41598-022-26499-y
9	Alejandro Barredo Arrieta et al.	2020	Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI	XAI Taxonomy	The study reports the application of Taxonomía XAI in the context described by its research objective.	https://doi.org/10.1016/j.inffus.2019.12.012
10	Pantelis Linardatos; Vasilis Papastefanopoulos; Sotiris Kotsiantis	2020	Explainable AI: A Review of Machine Learning Interpretability Methods	XAI Methods	The study reports the application of Métodos XAI in the context described by its research objective.	https://doi.org/10.3390/e23010018
11	Plamen, P. Angelov, et al.	2021	Explainable artificial intelligence: an analytical review	XAI	The study reports the application of XAI in the context described by its research objective.	https://doi.org/10.1002/widm.1424

No.	Authors	Year	Study	Method	Results	DOI
12	Dang Minh; H. Wang; Y.F. Li; T.N. Nguyen	2021	Explainable artificial intelligence : a comprehen sive review	XAI	The study reports the application of XAI in the context described by its research objective.	https://doi.org/10.1007/s10462-021-10088-y
13	Xiao Bai; Xiang Wang; Xianglong Liu; Qiang Liu; Jingkuan Song; Nicu Sebe; Been Kim	2021	Explainable deep learning for efficient and robust pattern recognition: A survey of recent developme nts	Explainable DL	The study reports the application of Explainable DL in the context described by its research objective.	https://doi.org/10.1016/j.patcog.2021.108102
14	Ding W, Abdel-Basset M, Hawash H, Ali AMA	2022	Explainabili ty of artificial intelligence methods, applications and challenges: A comprehen sive survey	XAI	The study reports the application of XAI in the context described by its research objective.	https://doi.org/10.1016/j.ins.2022.10.013
15	W. James Murdoch; Chandan Singh; Karl Kumbier; Reza Abbasi-Asl; Bin Yu	2019	Definitions, methods, and applications in interpretabl e machine learning	Interpretabi lity	The study reports the application of Interpretabi lidad in the context described by its research objective.	https://doi.org/10.1073/pnas.1900654116

No.	Authors	Year	Study	Method	Results	DOI
16	Joaquín Borrego-Díaz; Juan Galán-Páez	2022	Explainable Artificial Intelligence in Data Science	XAI socio-technical	The study reports the application of XAI socio-técnica in the context described by its research objective.	https://doi.org/10.1007/s11023-022-09603-z
17	Meske C, Bunde E, Schneider J, Gersch M	2020	Explainable Artificial Intelligence : Objectives, Stakeholders, and Future Research Opportunities	Objectives/ stakeholders	The study reports the application of Objetivos/stakeholders in the context described by its research objective.	https://doi.org/10.1080/10580530.2020.1849465
18	Fabien Charmet et al.	2022	Explainable artificial intelligence for cybersecurity: a literature survey	XAI + Cybersecurity	The study reports the application of XAI + ciberseguridad in the context described by its research objective.	https://doi.org/10.1007/s12243-022-00926-7

No.	Authors	Year	Study	Method	Results	DOI
19	Haque AKMB, Islam AKMN, Mikalef P	2022	Explainable Artificial Intelligence (XAI) from a user perspective : A synthesis of prior literature and problematizing avenues for future research	XAI user	The study reports the application of XAI usuario in the context described by its research objective.	https://doi.org/10.1016/j.techfore.2022.122120
20	Amiri SS, Mottahedi S, Lee EE, Hoque S	2021	Peeking inside the black-box: Explainable machine learning applied to household transportation energy consumption	XAI power consumption	The study reports the application of XAI consumo energético in the context described by its research objective.	https://doi.org/10.1016/j.compenvurbsys.2021.101647
21	Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR	2022	Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)	XAI Assessment	The study reports the application of Evaluación XAI in the context described by its research objective.	https://doi.org/10.1016/j.cmpb.2022.107161

No.	Authors	Year	Study	Method	Results	DOI
22	Di Martino F, Delmastro F	2022	Explainable AI for clinical and remote health applications : a survey on tabular and time series data	XAI Time Series	The study reports the application of XAI series temporales in the context described by its research objective.	https://doi.org/10.1007/s10462-022-10304-3
23	Payrovnazir i SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, Liu X, He Z	2020	Explainable AI models using real-world data: methodological lessons for critical domains	XAI Assessment	The study reports the application of Evaluación XAI in the context described by its research objective.	https://doi.org/10.1093/jamia/ocaa053
24	Miruna-Adriana Clinciu; Helen Hastie	2019	A Survey of Explainable AI Terminology	XAI Terminology	The study reports the application of Terminología XAI in the context described by its research objective.	https://doi.org/10.18653/v1/W19-8403
25	David S. Watson; Luciano Floridi	2021	The explanation game: a formal framework for interpretable machine learning	XAI Formal Framework	The study reports the application of Marco formal XAI in the context described by its research objective.	https://doi.org/10.1007/s11229-020-02629-9

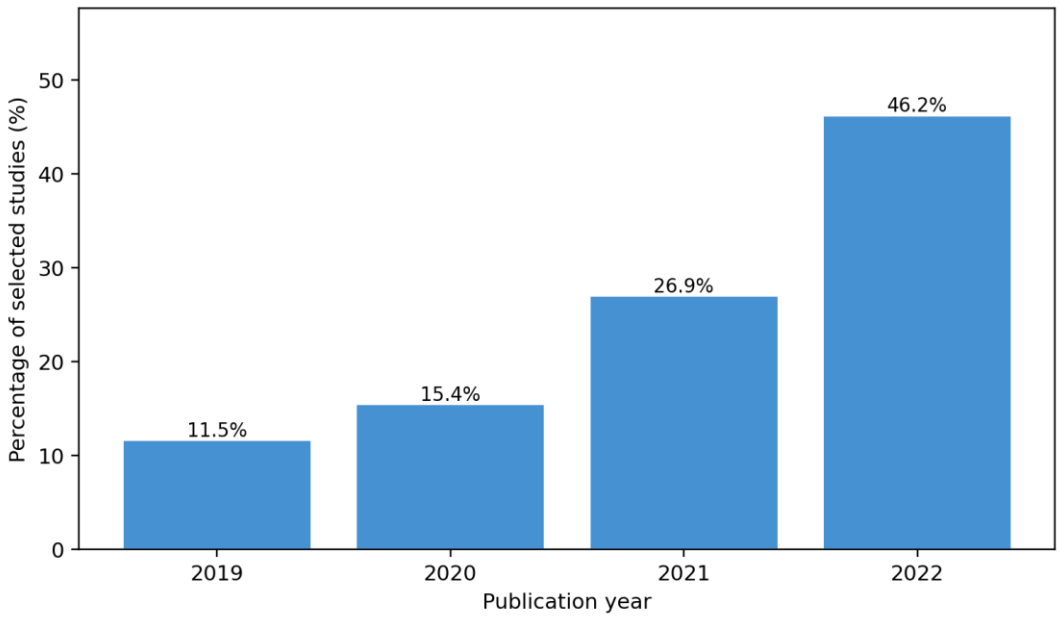
No.	Authors	Year	Study	Method	Results	DOI
26	Tim Miller	2019	Explanation in artificial intelligence : Insights from the social sciences	Human-centered explanation	The study reports the application of Explicación centrada en humanos in the context described by its research objective.	https://doi.org/10.1016/j.artint.2018.07.007

Note. The matrix reproduces the final selected corpus documented in the evidence workbook. DOI links are retained for verification.

Table 1 confirms that the final evidence base contains 26 individually identifiable studies with DOI. The matrix also shows that the corpus is methodologically mixed rather than dominated by a single design. This diversity is analytically useful because it combines conceptual consolidation with application-oriented evidence, but it also prevents treating all reported outcomes as directly comparable. The study type and technique columns therefore function as interpretive qualifiers: conclusions are drawn within compatible groups and are not converted into a single pooled effectiveness estimate.

Figure 2.

Percentage distribution of selected XAI studies by publication year



Note. Percentages were calculated from the 26 selected studies in the evidence matrix; each study contributes once according to publication year.

Figure 2 shows a clear temporal concentration of the selected XAI literature toward the end of the review window. The largest share corresponds to 2022, with 12 of 26 studies (46.2%). This pattern is consistent with the rapid consolidation of explainability as a research concern after broader debates on interpretable machine learning had already established terminology and taxonomies. The result should be interpreted as the distribution of this selected corpus rather than as a bibliometric estimate of the entire field, but it nevertheless illustrates the increasing presence of application-oriented XAI studies in energy and electrical contexts during 2021–2022.

DISCUSSION

The results indicate that explainability acquires meaning only when it is related to a specific electrical task. An explanation for load forecasting answers questions different from one used in photovoltaic control or stability analysis. The corpus combines interpretable models and post hoc techniques, but its usefulness cannot be judged solely by the existence of a visualization or a ranking of variables. It is necessary to identify which decision it intends to support, who receives the explanation, and which error would be relevant. This functional orientation makes it possible to distinguish technical transparency from operational utility. An explanation may be mathematically consistent and, at the same time, insufficient for an operator if it does not express understandable relationships within the physical context of the system. (1-5)

Post hoc techniques expand the possibility of examining complex models without replacing them with simpler approximations. Feature attributions and local explanations allow us to explore which variables influenced a specific prediction, but introduce questions about stability and fidelity. Two explanations can change in the face of small perturbations even if the output of the model hardly varies. In electrical applications, this sensitivity must be known before using the explanation to support a decision. The studies reviewed show increasing interest in these methods, although the evaluation of explanatory quality is less uniform than that of predictive performance. Reporting consistency, sensitivity, and scope of each explanation would help to better interpret its practical value. (3,9-14)

Intrinsically interpretable models offer a different path: they incorporate structures whose logic can be directly examined. Transparent rules, trees, or formulations can facilitate traceability, although their desirability depends on the complexity of the problem and the relationship between accuracy and understanding. The review does not allow us to argue that one type of model is superior in all scenarios. In some applications, a simpler structure may be sufficient and provide greater clarity; in others, dynamics require more flexible models accompanied by additional explanations. The choice should be based on the risk of the decision and the needs of the user. Comparing performance and comprehensibility within the same task would provide more useful evidence than contrasting algorithms in different domains. (5,10-16)

Forecasting is an area where XAI can provide additional information to predictive error. Knowing which variables drive an estimate of load, wind, or PV generation allows you to examine consistency with expected relationships and detect unexpected dependencies. However, a plausible explanation does not demonstrate that the model has learned a causal relationship. The work in the corpus should

be interpreted as analysis of predictive behavior, unless the design allows for stronger inferences. This distinction is essential to prevent an attribution from becoming causal justification. The evaluation should combine accuracy, stability of explanation, and expert review when the output is to be incorporated into operational or planning decisions. (2,4-8)

When XAI is incorporated into control functions, the requirement changes because the explanation accompanies a potential action on the system. The case of reactive power control illustrates the need to relate recommendation, influencing variables, and physical constraints. A useful explanation should help to recognize why the model proposes an action and when that recommendation may no longer be reliable. The literature reviewed shows that this level of integration still requires evaluation criteria closer to operational use. It is not enough to visualize the importance of variables; it is worth studying whether the explanation improves supervision, error detection, or human intervention capacity. These questions connect interpretability with security and technical responsibility. (3,17-19)

Network stability analysis offers a particularly sensitive scenario for explainability. Decisions associated with frequency, security, or contingencies require understanding not only the prediction, but also the conditions under which it may fail. Studies that apply XAI to these problems show the potential to identify relevant variables and explore patterns that a complex model uses. Still, operational confidence requires verifying that these explanations are consistent in the face of rare perturbations and states. An unstable explanation can induce a misinterpretation even when the average accuracy is high. Therefore, predictive robustness and explanatory robustness should be evaluated together in applications where the cost of an incorrect decision is significant. (1,2,15)

The general literature of XAI reminds us that an explanation does not exist independently of its recipient. Engineers, operators, developers, and managers may need different levels of detail. The electrical corpus incorporates this issue unevenly: many works describe the explanatory method more accurately than the interaction with users. This asymmetry limits the possibility of affirming that a technique is understandable under real conditions. A user-centered evaluation should define task, prior knowledge, format, and expected decision, and measure whether the explanation improves understanding or error detection. Incorporating these tests would allow progress from interpretability calculated by the algorithm to explainability effectively useful for technical work. (17-19,24-26)

The absence of standardized metrics for explanatory quality makes it difficult to compare methods. Fidelity, stability, simplicity, and comprehensibility describe different properties and can come under stress. A very detailed explanation may represent the model better, but be difficult to use; a brief explanation may be clear and omit important relationships. The studies reviewed suggest that the evaluation should be multidimensional and linked to the objective. In electrical applications, it is advisable to add coherence with physical constraints and the ability to signal anomalous conditions. Reporting multiple properties avoids reducing explainability to a single score. It also makes it easier to identify which technique is suitable for diagnosis, auditing, decision support, or communication with users who have different levels of knowledge. (9-18)

https://doi.org/10.1016/j.egyai.2022.100169

The robustness of an explanation deserves a separate treatment from the robustness of the predictor. If small variations in data produce large changes in the reasons given, the user may receive contradictory messages in almost equivalent situations. This problem is relevant in electrical time series, where noise, seasonality, and operational changes are part of the environment. Future work should test stability against plausible perturbations and document when an explanation becomes unreliable. It would also be useful to test techniques on the same model and dataset to determine whether the conclusions depend on the explanatory method. Such tests would strengthen confidence without assuming that a visually intuitive output is necessarily stable or faithful. (10-14,20)

An important conceptual issue is to separate the explanation of the model and the explanation of the physical phenomenon. XAI techniques usually describe how a prediction relates to inputs based on the behavior learned by the algorithm. That result does not automatically equate to identifying causes of the electrical system. Confusing the two levels can lead to interpreting correlations as physical mechanisms. The review underscores the need to formulate precisely what is being explained: the output of the model, an association in the data, or a causal process supported by additional knowledge. Integrating physical constraints, experiments, or domain analysis can strengthen interpretation, but the statements must correspond to the design used. This distinction improves rigor and avoids overextending the results. (9,15,16,25)

The reproducibility of XAI depends on documenting both the predictive model and the explanation procedure. Variable selection, preprocessing, algorithm version, parameters, and sample used can modify the explanatory results. When these elements are not reported, it is difficult to determine why two studies obtain different patterns. The corpus analyzed shows sufficient diversity to justify more comprehensive reporting practices. Sharing code and data when possible would facilitate comparisons, but even without open access configurations and limitations can be described in detail. This transparency would allow methodological differences to be distinguished from differences specific to the electrical domain and would favor a more robust accumulation of evidence on which explanations are stable, understandable, and operationally relevant. (11-18)

The involvement of domain experts is a necessary step in evaluating utility beyond algorithmic metrics. An engineer can detect that a seemingly coherent explanation contradicts physical constraints or ignores variables essential to the operation. User testing also allows us to observe interpretation time, calibrated confidence, and the ability to recognize model errors. The literature reviewed points to this need, but human validations are still less frequent than technical demonstrations. Experimental designs with defined tasks and clearly characterized user groups would provide evidence on when XAI actually improves decision-making. Such information is especially relevant before incorporating explanations into interfaces used in supervision, control, or planning. (17-19,24-26)

The 26 included studies show that the next stage of XAI in electrical systems requires integrating accuracy, explainability, and usage conditions. Progress should not be measured just by incorporating a new interpretation technique into an existing predictor. Explicit explanatory questions, metrics commensurate with the recipient, robustness testing, and validation with domain knowledge are

https://doi.org/10.1016/j.egyai.2022.100169

needed. It is also a good idea to distinguish local explanations from global conclusions and avoid causal attributions that are not supported by the design. A cumulative agenda should compare methods on common electrical tasks and report limitations transparently. In this way, XAI could evolve from auxiliary visualizations to reliable tools for auditing, diagnosis, and decision support. (19-26)

CONCLUSIONS

The systematic review made it possible to respond to the general objective by integrating the evidence from 26 studies on explainable artificial intelligence (XAI) in electricity systems within the period 2018–2022. The results show an expanding field whose evolution combines conceptual consolidation and increasingly specific applications. The main conclusion is that the usefulness of the solutions does not depend only on the novelty of a technique, but also on its correspondence with specific operational needs, the quality of the information used and the possibility of being consistently incorporated into electrical infrastructures that have their own technical and organizational constraints.

The corpus shows that technological integration is a central criterion for interpreting the results of Explainable Artificial Intelligence (XAI) in electrical systems. Interoperability, data availability, scalability, security, maintainability, and compatibility with existing systems recur as conditions that can favor or limit the observed performance. Therefore, the positive results obtained in particular scenarios should not be automatically generalized to other networks, assets, or contexts. The future evaluation should incorporate these conditions as part of the performance and not treat them as separate secondary aspects of the main solution.

The methodological heterogeneity identified explains why it is not appropriate to establish a universal ranking of methods or to calculate an aggregate effect from results that respond to different objectives and metrics. The review provides greater value by organizing convergences, contrasts, and gaps within comparable categories. This conclusion guides the reading of the tables and figures: the frequencies describe the composition of the corpus and the matrices allow individual evidence to be traced, while the thematic interpretation relates the findings to the research problem without attributing equivalence to designs that were not designed to answer the same question.

Subsequent research should strengthen reproducibility through complete descriptions of data, configurations, procedures, and metrics, as well as incorporating comparisons under representative conditions and predefined evaluation criteria. It is also necessary to expand validation outside of controlled scenarios and clearly document transfer constraints. More comparable evidence would make it possible to determine which solutions maintain their performance in the face of changes in scale, topology, load, data quality, or communications infrastructure, and would make it easier for academic results to be translated into engineering decisions with less uncertainty.

Finally, the synthesis confirms that the development of explainable artificial intelligence (XAI) in electricity systems should be understood as a multidimensional process in which technology, data, architecture, and operational context converge. The selected studies offer a sufficient basis to identify consolidated lines and persistent challenges, but also show the need for more comprehensive

evaluations. The future agenda should prioritize verifiable, secure, interoperable, and technically justified solutions, with indicators linked to the actual purpose of use. This orientation would allow progress from feasibility demonstrations to cumulative evidence capable of supporting responsible adoptions in electricity systems.

BIBLIOGRAPHIC REFERENCES

1. Ram Machlev, Leena Heistrene, M. Perl, Kfir Y. Levy, Juri Belikov, Shie Mannor, Yoash Levron. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. Energy AI. 2022;9:100169. https://doi.org/10.1016/j.egyai.2022.100169.
2. Johannes Kruse, Benjamin Schäfer, Dirk Witthaut. Revealing drivers and risks for power grid frequency stability with explainable AI. Patterns. 2021;2(11):100365. https://doi.org/10.1016/j.patter.2021.100365.
3. Utama C, Meske C, Schneider J, Ulbrich C. Reactive power control in photovoltaic systems through (explainable) artificial intelligence. Appl Energy. 2022;328:120004. https://doi.org/10.1016/j.apenergy.2022.120004.
4. Haoqi Huang, Yue Hong, Huaizhi Wang. Probabilistic Prediction Intervals of Wind Speed Based on Explainable Neural Network. Front Energy Res. 2022;10:https://doi. org/https://doi.org/10.3389/fenrg.2022.934935.
5. Jihoon Moon, Sungwoo Park, Seungmin Rho, Eenjun Hwang. Interpretable Short-Term Electrical Load Forecasting Scheme Using Cubist. Comput Intell Neurosci. 2022;2022(1):6892995. https://doi.org/10.1155/2022/6892995.
6. Peng Han, Wei He, You Cao, YingMei Li, YunYi Zhang. Deep belief rule based photovoltaic power forecasting method with interpretability. Sci Rep. 2022;12(1):https://doi. org/https://doi.org/10.1038/s41598-022-18820-6.
7. Chakraborty D, Alam A, Chaudhuri S, Basagaoglu H, Sulbaran T, Langar S. Scenario-based prediction of climate change impacts on building cooling energy consumption with explainable artificial intelligence. Appl Energy. 2021;291:116807. https://doi.org/10.1016/j.apenergy.2021.116807.
8. Abdul Khalique Shaikh, Amril Nazir, Imran Khan, Abdul Salam Shah. Short term energy consumption forecasting using neural basis expansion analysis for interpretable time series. Sci Rep. 2022;12(1):https://doi. org/https://doi.org/10.1038/s41598-022-26499-y.
9. Alejandro Barredo Arrieta et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82-115. https://doi.org/10.1016/j.inffus.2019.12.012.
10. Pantelis Linardatos, Vasilis Papastefanopoulos, Sotiris Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy. 2021;23(1):18. https://doi.org/10.3390/e23010018.
11. Plamen P. Angelov et al. Explainable artificial intelligence: an analytical review. WIREs Data Min Knowl Discov. 2021;11(5):e1424. https://doi.org/10.1002/widm.1424.
12. Dang Minh, H. Wang, Y.F. Li, T.N. Nguyen. Explainable artificial intelligence: a comprehensive review. Artif Intell Rev. 2022;55(8):6341-6400. https://doi.org/10.1007/s10462-021-10088-y.

13. Xiao Bai, Xiang Wang, Xianglong Liu, Qiang Liu, Jingkuan Song, Nicu Sebe, Been Kim. Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. Pattern Recognit. 2021;120:108102. https://doi.org/10.1016/j.patcog.2021.108102.
14. Ding W, Abdel-Basset M, Hawash H, Ali AMA. Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. Inf Sci. 2022;615:238-292. https://doi.org/10.1016/j.ins.2022.10.013.
15. W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, Bin Yu. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci U S A. 2019;116(44):22071-22080. https://doi.org/10.1073/pnas.1900654116.
16. Joaquín Borrego-Díaz, Juan Galán-Páez. Explainable Artificial Intelligence in Data Science. Minds Mach. 2022;32(3):485-531. https://doi.org/10.1007/s11023-022-09603-z.
17. Meske C, Bunde E, Schneider J, Gersch M. Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. Inf Syst Manag. 2022;39(1):53-63. https://doi.org/10.1080/10580530.2020.1849465.
18. Fabien Charmet et al. Explainable artificial intelligence for cybersecurity: a literature survey. Ann Telecommun. 2022;77(11-12):789-812. https://doi.org/10.1007/s12243-022-00926-7.
19. Haque AKMB, Islam AKMN, Mikalef P. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. Technol Forecast Soc Change. 2023;186:122120. https://doi.org/10.1016/j.techfore.2022.122120.
20. Amiri SS, Mottahedi S, Lee EE, Hoque S. Peeking inside the black-box: Explainable machine learning applied to household transportation energy consumption. Comput Environ Urban Syst. 2021;88:101647. https://doi.org/10.1016/j.compenvurbsys.2021.101647.
21. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011-2022). Comput Methods Programs Biomed. 2022;226:107161. https://doi.org/10.1016/j.cmpb.2022.107161.
22. Di Martino F, Delmastro F. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. Artif Intell Rev. 2023;56(6):5261-5315. https://doi.org/10.1007/s10462-022-10304-3.
23. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, Liu X, He Z. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J Am Med Inform Assoc. 2020;27(7):1173-1185. https://doi.org/10.1093/jamia/ocaa053.
24. Miruna-Adriana Clinciu, Helen Hastie. A Survey of Explainable AI Terminology. Proceedings of NL4XAI 2019 / ACL. 2019. https://doi.org/10.18653/v1/W19-8403
25. David S. Watson, Luciano Floridi. The explanation game: a formal framework for interpretable machine learning. Synthese. 2021;198(10):9211-9242. https://doi.org/10.1007/s11229-020-02629-9.
26. Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell. 2019;267:1-38. https://doi.org/10.1016/j.artint.2018.07.007.

ETHICAL CONSIDERATIONS

Not applicable. This study is a review of published literature and did not involve direct participation of human beings, animals, interventions, or collection of identifiable personal data.

ARTIFICIAL INTELLIGENCE (AI) USE STATEMENT

The authors declare that the generative artificial intelligence tool ChatGPT (OpenAI) was used exclusively to support the organization of the article and translation of text. Its use did not determine the scientific content, study design, source selection, data analysis, interpretation of results, or conclusions. Full responsibility for the content remains with the authors.

FUNDING

This research did not receive any financial support.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTION

Author Names and Initials: Danner Anderson Figueroa Guerra (DAFG)

1. Conceptualization: DAFG
2. Data curation: DAFG
3. Formal analysis: DAFG
4. Acquisition of financing: not applicable
5. Investigation: DAFG
6. Methodology: DAFG
7. Project administration: DAFG
8. Resources: DAFG
9. Software: not applicable
10. Supervision: DAFG
11. Validation: DAFG
12. View: DAFG
13. Original draft writing: DAFG
14. Writing, review and editing: DAFG