Explainable AI in Healthcare: Systematic Review of Clinical Decision Support Systems

NOOR A. AZIZ[1, 2], AWAIS MANZOOR[1, 2, 3], MUHAMMAD DEEDAHWAR MAZHAR QURESHI[2,4], M. ATIF QURESHI[1, 2, 4], WAEL RASHWAN.[2]

[1]ADAPT Centre
[2]eXplainable Analytics Group, Faculty of Bussiness, Technological University of Dublin
[3]Artificial Intelligence and Cognitive Load Research Lab, Technological University Dublin.
[4]Science Foundation of Ireland, Centre for Research Training in Machine Learning (ML-Labs), Technological University Dublin.

ABSTRACT

This systematic review examines the evolution and current landscape of eXplainable Artificial Intelligence (XAI) in Clinical Decision Support Systems (CDSS), highlighting significant advancements and identifying persistent challenges. Utilising the PRISMA protocol, we searched major indexed databases such as Scopus, Web of Science, PubMed, and the Cochrane Library, to analyse publications from January 2000 to April 2024. This timeframe captures the progressive integration of XAI in CDSS, offering a historical and technological overview. The review covers the datasets, application areas, machine learning models, explainable AI methods, and evaluation strategies for multiple XAI methods.

Analysing 68 articles, we uncover valuable insights into the strengths and limitations of current XAI approaches, revealing significant research gaps and providing actionable recommendations. We emphasise the need for more public datasets, advanced data treatment methods, comprehensive evaluations of XAI methods, and interdisciplinary collaboration. Our findings stress the importance of balancing model performance with explainability and enhancing the usability of XAI tools for medical practitioners. This research provides a valuable resource for healthcare professionals, researchers, and policymakers seeking to develop and evaluate effective, ethical decision-support systems in clinical settings.

INDEX TERMS Clinical Decision Support Systems, eXplainable Artificial Intelligence, Machine Learning, Computer Aided Diagnosis, Electronic Health Record

I. INTRODUCTION

THe advancements in computer science, particularly in Machine Learning, in the 21st century have been pivotal in ushering in the fourth industrial revolution. This has given rise to an interdisciplinary nexus of technological applications that address challenges beyond the traditional boundaries of computer science. One significant impact area is healthcare, where these technologies have seen increased acceptance and adoption for mission-critical tasks.

Integrating AI and ML technologies in healthcare has enhanced decision-making capabilities, yet it has also underscored the need for greater transparency in these systems. The COVID-19 pandemic, in particular, highlighted the importance of such technologies in managing new and evolving medical knowledge during critical situations. AI and ML systems have been crucial in areas such as disease classification and cancer diagnosis. However, the accuracy and precision of these systems, while important, are not sufficient on their own. It is essential to empower decision-makers with interpretable methods that allow stakeholders to understand the technological decisions, evaluate their merits, and make necessary adjustments. Rapid advancements in XAI have made model explainability a central concern, serving as a much-needed bridge between AI and its application domains, including healthcare, social sciences, and engineering.

This need for explainability extends naturally to clinical decision-making for healthcare professionals, particularly in fields that already rely heavily on technology, such as ML/AI models integrated into standard protocols. As healthcare continues to embrace digital transformation, implementing explainable AI within Clinical Decision Support Systems (CDSS) is crucial for ensuring informed, ethical, and effective patient care.

TABLE 1: List of abbreviations and their descriptions.

Abbreviation	Description	Abbreviation	Description
RNN	Recurrent Neural Network	GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory	CNN	Convolutional Neural Network
LR	Logistic Regression	SVM	Support Vector Machine
RF	Random Forest	XGB	Extreme Gradient Boost
MLP	Multi-Layer Perceptron	ANN	Artificial Neural Network
DT	Decision Trees	GBC	Gradient Boosting Classifier
GNB	Gaussian Naïve Bayes	KNN	k-Nearest Neighbors
LDA	Linear Discriminant Analysis	LGBM	Light Gradient Boosting Machine
DNN	Deep Neural Network	CART	Classification and Regression Trees
NBM	Naive Bayes Multinomial	GLM	Generalised Linear Model
EN	Elastic Net	TB	Tree Boosting
ICE	Individual Conditional Expectation	GBDT	Gradient Boosted Decision Trees
FCN	Fully Convolutional Neural Network	GB	Gradient Boosting
SVC	Support Vector Classification	ETC	Extra Trees Classifier
SGD	Stochastic Gradient Descent	LGB	Light Gradient Boosting Machine
CDSS	Clinical Decision Support System	EMR	Electronic Medical Records
EHR	Electronic Health Records	LGA	Large for Gestational Age
LIME	Local Interpretable Model-agnostic Explanations	AGRAD	Attention Gradient
LORE	LOcal Rule-based Explanations	LRP	Layer-wise Relevance Propagation
GradCAM	Gradient-weighted Class Activation Mapping	PDP	Partial dependence plots
SHAP	Shapley Additive Explanations	CIU	Contextual Importance and Utility
FI	Feature Importance	BoCSoR	Boundary Crossing Solo Ratio
PFI	Permutation Feature Importance	ELI5	Explain Like I am 5
ABELE	Anchors Basic Explanation Linked-Examples Enhanced Explanation		

A. NEED FOR EXPLAINABLE CDSS IN HEALTHCARE

Clinical Decision Support Systems (CDSS) are essential tools in modern healthcare, aiding doctors, nurses, and pharmacists in making better-informed decisions about patient care [1]. These systems range from basic information providers, such as medication and lab result databases like pharmacy information systems [2], to advanced algorithms that recommend personalised treatments [3]. Some CDSSs operate automatically, providing instant guidance to healthcare professionals [4], while others require manual input, such as clinical guidelines [5].

The digital transformation in healthcare has led to exponential growth in data generation, particularly through the adoption of Electronic Health Records (EHRs). EHRs chronicle comprehensive patient health information, offering a digital alternative to traditional paper records. Leveraging this data, advanced CDSS has been developed to assist healthcare providers in making informed decisions based on patient-specific conditions.

The potential of CDSS to reduce errors in decision-making and improve patient outcomes is well-documented. However, the mere presence of these systems does not automatically guarantee better patient care. While studies have shown a significant reduction in decision-making errors [6], the effectiveness of CDSS is highly dependent on the quality of the systems and the accuracy of the data they process.

The utility of CDSS becomes particularly apparent in scenarios requiring swift and informed decision-making. By providing clinicians with immediate access to relevant patient data and the latest medical knowledge, CDSS empowers healthcare professionals to navigate complex and voluminous data, ensuring that critical decisions regarding patient care are both informed and timely. The ultimate goal of deploying CDSS in healthcare is to harness the best available evidence and insights, thereby enhancing patient care outcomes through informed clinical decision-making.

XAI is crucial in this context, as it addresses the need for transparency and interpretability in CDSS. Integrating XAI into CDSS ensures that the decisions made by these systems are not only accurate but also understandable to clinicians. This transparency is essential for gaining the trust of healthcare providers and ensuring that the systems' recommendations can be effectively scrutinized and validated.

B. MOTIVATION FOR THE WORK

The integration of XAI into CDSS represents a significant advancement in healthcare technology. Despite various review studies highlighting the multifaceted challenges and opportunities of this integration, there is a need for a comprehensive evaluation of current research directions and future goals.

Several studies have underscored the potential and necessity of XAI in enhancing the transparency and reliability of CDSS. For instance, Antoniadi *et al.* [7] showcased successful integrations of XAI, providing foundational insights into its application within CDSS. Similarly, Du *et al.* [8] emphasized the critical importance of explainability in CDSS, particularly in specific healthcare domains such as pregnancy care, where inclusivity and extensive validation are essential. Vasey *et al.* [9] revealed a lack of sufficient evidence to conclusively assert that ML-based CDSS enhances physician diagnostic capabilities, highlighting limitations such as small participant sizes, biases, and inadequate consideration of human factors. Wang *et al.* [10] identified several obstacles to the effective implementation of XAI in CDSS, including

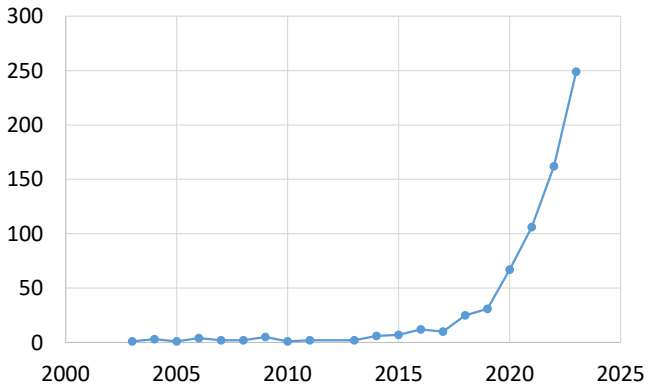


FIGURE 1: Search results from all databases

technical, process-related, attitudinal, informational, usability, and environmental challenges. Similarly, Moazemi *et al.* [11] similarly reported issues with validation and interoperability in CDSS. Xu *et al.* [12] suggested that future studies should focus on developing a formalism for defining interpretability, identifying its properties, and creating appropriate and objective metrics for evaluating interpretability.

The analysis of recent systematic literature reviews on XAI-based CDSS indicates an inadequacy in studies that holistically address critical aspects of these systems (see Table 2). This observation underscores the need for a systematic review that maps the current research landscape, identifies existing limitations, and guides future directions. By addressing this gap, researchers and practitioners can better understand the challenges and opportunities presented by integrating XAI into CDSS, facilitating the development of more effective, ethical, and user-friendly AI-driven healthcare solutions.

This work aims to provide a thorough review of the current state of XAI in CDSS, evaluating its implementation, challenges, and potential for future research. It seeks to offer valuable insights for healthcare professionals, researchers, and policymakers interested in developing and evaluating transparent and reliable AI-based decision support systems in clinical settings.

C. RESEARCH QUESTIONS

This literature review examines the current use of eXplainable AI methods in AI-assisted healthcare systems. While AI-powered healthcare systems have been extensively studied, the application of XAI to this domain is still emerging. Healthcare is rapidly becoming a significant area for XAI applications. This review aims to assess the adoption of XAI across different healthcare domains and its implications. Additionally, it explores the interactions between various ML models and XAI methods regarding their compatibility, necessity, and effectiveness in enhancing interpretability. These queries are formalised into the following research questions:

1) Which specific healthcare domains have extensively adopted XAI methods?
2) What types of datasets are commonly used in explainable CDSS, and what are their key characteristics?
3) What are the prevailing trends and most effective machine learning models utilised in explainable healthcare systems?
4) What are the most commonly used XAI methods in electronic healthcare, and how do they contribute to the interpretability of machine learning models?
5) How do multiple XAI methods contribute to ensuring explainability in the healthcare decision-making process?

II. METHODOLOGY

To conduct this systematic review, we adhered to rigorous methodology, drawing upon the established literature review framework [18] and further aligning our review protocols with PRISMA-P guidelines [19] to ensure a comprehensive and methodological approach. Figure 2 illustrates the detailed implementation of the PRISMA-P protocol tailored for our review objectives.

The following sections discuss the specifics of the SLR methodological approach, including eligibility criteria, identification of relevant information sources, formulation of the search strategy, and the screening process undertaken to curate the final set of articles for analysis.

A. ELIGIBILITY CRITERIA

This review focuses on the applications of XAI in healthcare, particularly healthcare/clinical decision support systems (CDSS). The purpose is to assess the overall progress of artificial intelligence in the last two decades and the gradual acceptance and need for XAI to make AI-based systems more reliable in critical domains such as healthcare. It also examines the challenges artificial intelligence has encountered during this period and the current obstacles to adopting more XAI-based systems. Additionally, it aims to identify future research insights in this specific domain.

The study also aims to determine the types of data supported by artificial intelligence-based clinical decision support systems, the implementation of the systems, and the outcome of the systems. The terms that describe our work most accurately are "clinical decision support systems" and "explainable artificial intelligence." These terms help us find all the related studies in scientific databases, and all the search terms and search strings were designed based on these two terms.

The review included studies that discuss CDSS regardless of the type of disease. However, the type of disease was still considered a significant parameter for the classification of studies and the data type they used. Studies that did not discuss the technical aspects of CDSS, philosophical studies, review studies, conference articles, and discussion studies were excluded from the review.

Furthermore, the research studies were selected based on their use of explainability methods, i.e., at least one XAI

TABLE 2: Summary of CDSS Literature

Ref	Published	Duration	Application Area	ML Models	XAI Methods	Multiple XAI Methods	Dataset Characteristics	Remarks
[7]	2021	2008-2020	-	-	Low	-	-	Touches upon XAI methods in CDSS without focusing on specific ML models.
[13]	2022	2011 - 2022	Low	Low	High	-	-	Focuses on XAI methods with limited discussion on ML models and application areas.
[12]	2023	2011-2020	-	-	Low	-	-	Discusses CDSS from a technological and medical perspective, with no detailed discussion of ML models, XAI methods, and datasets.
[14]	2023	2018 - 2022	Low	Low	Medium	-	Medium	Focuses on XAI methods and the datasets for CDSS, with a limited discussion on ML models and application areas and no discussion of using multiple XAI methods in a single study.
[15]	2023	by Oct 2022	-	-	High	-	-	Reviews the implementation of XAI methods and their challenges from a physician's perspective, with no discussion of ML models, datasets, and applications.
[16]	2023	2019-2022	Medium	Medium	High	-	Medium	Focuses on different XAI methods mostly on image-bases datasets, and lacks discussion on free-text and tabular datasets.
[17]	2023	2020 - 2022	-	Low	Medium	-	Low	Focuses on the taxonomy of XAI methods in the medical field, lacking discussion on application areas, with limited discussion on datasets and ML models.
This review	2024	2003-2023	High	High	High	High	High	Provides a comprehensive analysis of XAI methods applied to different ML model categories across various healthcare application areas, detailing dataset characteristics and lists tradeoffs between XAI methods used.

method (unlike EDA strategies) to explain at least one black-box or non-interpretable ML model to decision-makers. Specific XAI methods were not a requirement for inclusion in the review.

B. DATABASES

We utilised multiple databases to gather information, including Scopus, Web of Science, Cochrane Library, IEEE Explore, PubMed, and Science Direct. Although we received almost identical outcomes from different databases, we focused on four primary databases to eliminate redundancy and duplicate publications. These four databases are Scopus, Web of Science, PubMed, and Cochrane Library. Scopus and Web of Science are vast databases encompassing various research topics and publications, making them reliable sources for literature in any field. Meanwhile, PubMed and Cochrane Library are important sources of knowledge for medical and biomedical studies. PubMed covers the overall advancements in medical expertise, while Cochrane Library is a valuable source of clinical studies that examine all clinical-related practices, implementations, and clinical outcomes. Although we tested our search terms on IEEE Explore, we didn't achieve enough results. We also discovered that Science Direct produced nearly identical outcomes to Scopus and Web of Science. However, due to Scopus and Web of Science's broad range of multidisciplinary coverage, we focused on these two databases.

C. SEARCH STRATEGY

We considered two major terms that define our research topic. These terms are "clinical decision support systems" and "explainable artificial intelligence." We have designed our search strategy into three parts based on these two terms. The first part of the search strategy focuses on explainability, covering all the synonyms and explainability-related words and combining them with the binary OR operator. The second part covers artificial intelligence in publications related to artificial intelligence, while the third part represents clinical decision systems, covering all the synonyms and related words to clinical decision support systems. To make these three parts of the search term work together, we have combined them using the binary AND operator; see Table 3.

After designing the search terms, we selected the topic title, abstract, and keywords and applied the search string, and adapted the search query according to each database search policy (see Table 4). Also, we applied the duration of the search between 01-Jan-2000 and 31-Dec-2023 and collected a total of 1226 research articles[1]. Furthermore,

[1] 451, 351, 229, and 195 results from Scopus, Web of Science, Cochrane Library, and PubMed, respectively.

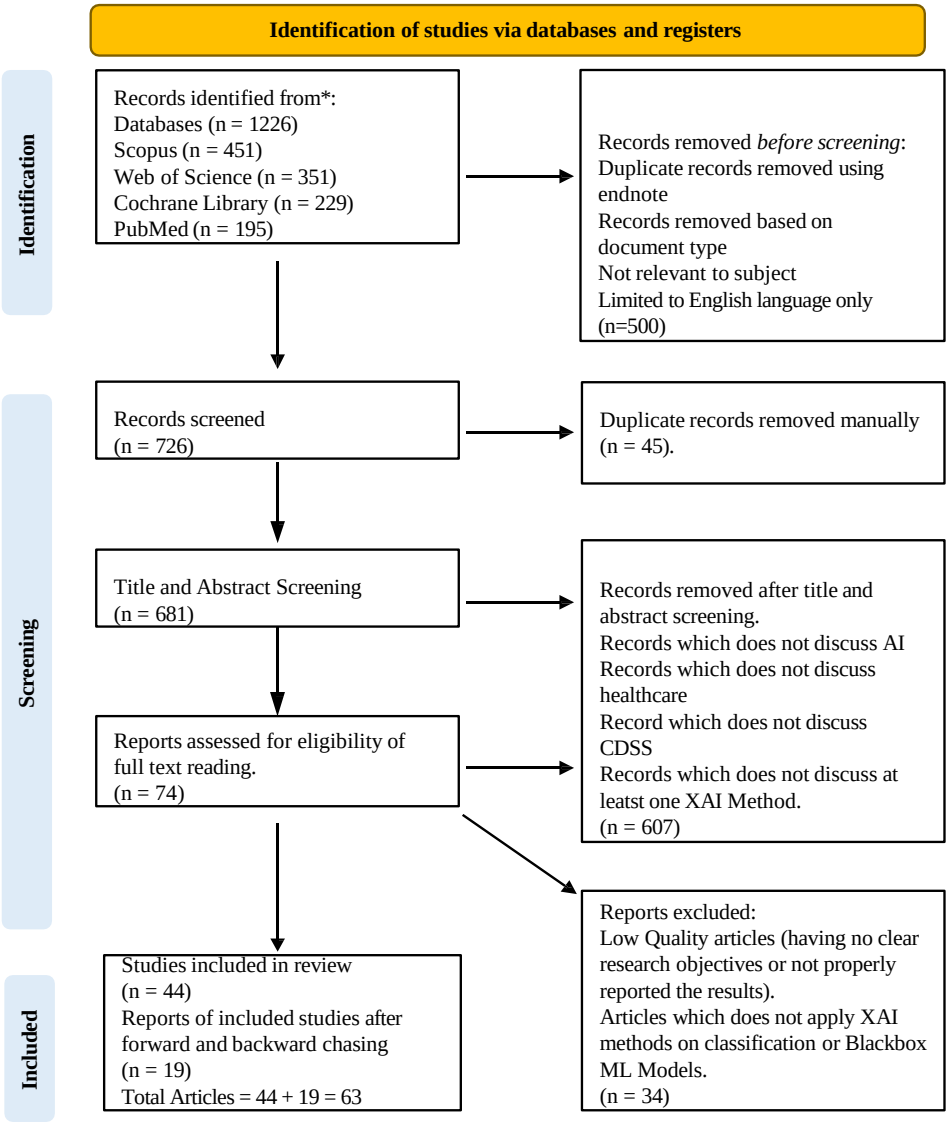


FIGURE 2: PRISMA Diagram

articles from 2024 are also included until April, totaling in 5 additional articles.

D. SCREENING OF STUDIES

We took several steps to ensure we only considered high-quality, relevant articles. Initially, we applied filters based on document type, subject matter, and language (English only) to limit our search results only to include peer-reviewed journal articles that focused on the topic of XAI and CDSS. After applying these filters, we used Endnote reference manager software to detect and remove duplicate articles, and we also removed a small number of duplicate articles through manual intervention. After this initial screening, we were left with 681 articles, which we reviewed based on the title and abstract to determine if they met our inclusion and exclusion criteria (see Table 5). Of these, 74 articles met our criteria and were subjected to a full-text review. During this review, we only included articles that have clear research objectives, properly reported the results, and discussed the application of the XAI method on classification or black-boxed machine learning models and at least one explanation method to clinical data. Ultimately, we were left with 44 articles that matched our inclusion criteria. To ensure we didn't miss any relevant articles, we also performed forward and backward reference chasing, which led us to include 19 additional articles that matched our criteria and 5 articles published until April 2024.

Search Terms	Operator	Search Terms	Operator	Search Terms
(explainable OR interpretable OR transparent OR accountable OR human-interpretable OR human-centered)	AND	(AI OR ML OR Deep Learning OR machine AND learning OR deep AND learning OR artificial AND intelligence OR computational AND intelligence) OR XAI)	AND	(((health* OR *medical OR clinical) AND decision* AND (support OR making OR aids) AND system*) OR cdss OR cds))

TABLE 3: Search Terms

Name of Database	Search String
Scopus	TITLE-ABS-KEY (((explainable OR interpretable OR transparent OR accountable OR human-interpretable OR human-centered) AND (ai OR ml OR DL OR machine AND learning OR deep AND learning OR artificial AND intelligence OR computational AND intelligence) OR XAI) AND (((health* OR *medical OR clinical) AND decision* AND (support OR making OR aids) AND system*) OR cdss OR cds))
Web of Science	AB=(((explainable OR interpretable OR transparent OR accountable OR human-interpretable OR human-centered) AND (ai OR ml OR DL OR machine AND learning OR deep AND learning OR artificial AND intelligence OR computational AND intelligence) OR XAI) AND (((health* OR medical OR clinical) AND decision* AND (support OR making OR aids) AND system*) OR cdss OR cds))
Cochrane Library	((Explainable OR Interpretable OR Transparent OR Accountable OR Human-interpretable OR Human-Centered) AND (AI OR ML OR DL OR machine learning OR Deep learning OR artificial intelligence OR computational intelligence) OR XAI) AND (((Health* OR *medical OR clinical) AND decision* AND (support OR making OR aids) AND system*) OR CDSS OR CDS)
PubMed	((("Explainable"[Title/Abstract] OR "interpretable"[Title/Abstract] OR "Transparent"[Title/Abstract] OR "Accountable"[Title/Abstract] OR "Human-interpretable"[Title/Abstract] OR "Human-Centered"[Title/Abstract]) AND ("AI"[Title/Abstract] OR "ML"[Title/Abstract] OR "DL"[Title/Abstract] OR "machine learning"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "artificial intelligence"[Title/Abstract] OR "computational intelligence"[Title/Abstract])) OR "XAI"[Title/Abstract]) AND ((("health*"[Title/Abstract] OR "medical"[Title/Abstract] OR "clinical"[Title/Abstract]) AND "decision*"[Title/Abstract] AND ("support"[Title/Abstract] OR "making"[Title/Abstract] OR "aids"[Title/Abstract]) AND "system*"[Title/Abstract]) OR "CDSS"[Title/Abstract] OR "CDS"[Title/Abstract])

TABLE 4: Summary of Database Searches

Inclusion	Exclusion
• Discuss XAI in healthcare • Discuss XAI and CDSS. • Articles utilising ML models and XAI methods with a clinical dataset.	• Articles lacking attention on healthcare and CDSS. • Articles not written in English • Duplicate articles • Articles involving fuzzy classifiers • Philosophical Studies • Review studies • Conference articles • Discussion studies

TABLE 5: Inclusion and Exclusion Criteria

E. DATA EXTRACTION AND STUDY CHARACTERISTICS

This review includes articles published from January 2000 to April 2024. After applying the search strings and filtering studies according to the inclusion and exclusion criteria, we were left with studies predominantly published in recent years, as shown in Figure 3. Specifically, seven articles were published in 2020, 9 in 2021, 14 in 2022, 33 in 2023, and 5 articles published by April 2024. The gradual increase in studies over recent years indicates growing interest from researchers and stakeholders in applying artificial intelligence within the healthcare domain. Consequently, it is an opportune time to evaluate the current state of XAI in CDSS, understand the existing challenges, and provide insights for future implementation and research.

The studies selected for this review were subjected to a thorough data extraction process to capture relevant characteristics. Key data extracted included study design, sample size, XAI methods used, ML models implemented, datasets characteristics, and the specific healthcare applications addressed. This detailed extraction aimed to ensure a comprehensive understanding of how XAI is being utilised within CDSS and to identify any prevalent trends or gaps in the research.

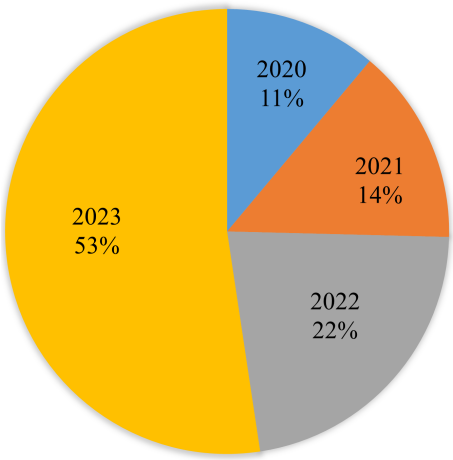


FIGURE 3: Studies published from 2020 to 2023 are included (*Statistics for 2024 are not shown as the year is ongoing).

III. APPLICATION AREAS AND DATASETS

The articles included in the review were analyzed and categorised based on their applications in specific disease types. The review identifies 19 application categories summarized in Table 6. The six most prominent categories are discussed in detail below, while the remaining categories with less frequent usage are combined into a final subsection. Detailed

descriptions of the datasets relevant to each application area are presented in Table 8 and 9, including their detailed descriptions.

TABLE 6: Overview of datasets by application areas, availability, and type of data

Application Area	Availability	Dataset Type	References
Neurological (17)	Pub (5)	Tab (3)	[20]–[22]
		Img (1)	[23]
		Txt (4)	[24]
	Prv (12)	Tab (11)	[25], [26], [27], [28], [29] , [30], [31], [22], [32], [21], [33]
		Img (1)	[34]
Cancer (13)	Pub (7)	Tab (4)	[21], [35]–[37]
		Img (3)	[38]–[40]
	Prv (6)	Tab (3)	[33], [41], [42]
		Img (3)	[43]–[45]
Cardiovascular (7)	Pub (3)	Tab (3)	[21], [46], [47]
	Prv (4)	Tab (4)	[48], [49], [50], [33]
Diabetes (6)	Pub (4)	Tab	[21], [51], [52]
	Prv (3)	Tab	[53]–[55]
COVID-19 (7)	Pub (3)	Tab (2)	[56], [57]
		Img (2)	[58]
	Prv (3)	Tab (3)	[59]–[61]
Mortality Risk (ICU) (4)	Prv (4)	Tab (4)	[62]–[65]
Endoscopy (3)	Pub (3)	Img (3)	[66]–[68]
Skin (2)	Pub (1)	Img (1)	[69]
	Prv (1)	Img (1)	[70], [71]
Anti-microbial (2)	Prv (2)	Tab (2)	[72], [73]
Pregnancy (3)	Pub (1)	Tab (1)	[74]
	Prv (2)	Tab (2)	[75], [76]
Cytokines (1)	Prv (1)	Tab (1)	[77]
Pneumonia (4)	Pub (3)	Img (3)	[58], [78]
	Prv (1)	Img (1)	[78]
Hepatitis (1)	Pub (1)	Tab (1)	[79]
Obesity (1)	Pub (1)	Tab (1)	[50]
Pulse Wave Classification (1)	Prv (1)	Img (1)	[80]
Surgical (1)	Prv (1)	Tab (1)	[46]
Acute Disease (1)	Prv (1)	Tab (1)	[81]
Chronic Disease (1)	Pub(1)	Tab (1)	[82]
Medical Abstracts (General)	Pub (1)	Text (1)	[83]

A. NEUROLOGICAL

Neurological conditions are the most studied category, with 17 studies focusing on diagnosis and assessment (see Table 6). Neurological conditions include, Alzheimer's disease [28], [30], [34], stroke [20], [22], [32], [84], dementia [25], Parkinson's disease [26], Cerebrovascular issues [23], brain injuries [27], depressive disorder [31], language behavior based mental health issues [24] and brain connectivity networks [29]. A total of 12 private datasets and five public datasets were used, featuring tabular (14), image (2), and text data (4). These datasets are often imbalanced and contain missing values (see Table 8), necessitating careful preprocessing to ensure robust analysis and model performance.

B. CANCER

Cancer detection and diagnosis constitute a significant portion of the reviewed studies, with 13 studies addressing various types of cancer, including cervical [35], liver [33], multiple myeloma [36], prostate [38], lung [33], [41], [85], glioma brain [39], [40], breast [21], [33], [44], nasopharyngeal [42], colorectal [43] and tumor [45]. Both image-based and tabular datasets were utilised, with a distribution of 6 image-based and seven tabular datasets, of which three are public in both categories while the remaining are private. Similar to neurological datasets, these datasets are frequently imbalanced and contain missing values (see Table 8 and 9), which must be addressed during preprocessing.

C. CARDIOVASCULAR

Seven studies focused on cardiovascular conditions such as cardiac arrest [49], heart failure with coronary heart disease [48], comorbidity [50], hypertensive heart disease [33], myocardial infarction [46], assessing conditions through electrocardiograms [47] and cardiography [21]. All datasets in this category are tabular, with three publicly available and four being private. These datasets are typically imbalanced and contain noise and missing values, impacting the choice of predictive models and preprocessing techniques.

D. COVID-19 AND DIABETES

Seven studies each focused on COVID-19 and diabetes. COVID-19 studies explored COVID-19 prediction [56], [86], ICU admissions [59], diagnosis from influenza-like illness [60], triage-prediction system [57], severity risk [61], severe community-acquired pneumonia and respiratory infections [58]. While Diabetes studies explored predicting large gestational age (LGA) in overweight and obese female patients [52]–[54], general classification and prediction of diabetes [51], and addressing diabetes retinopathy in type 2 diabetes patients [21], [55].

The datasets for both categories included all data in tabular format except two images in the COVID-19 Category. While three public and three private datasets existed in each category, both had missing values and outliers. The diabetes datasets showed less imbalance compared to COVID-19,

with one exception of a highly imbalanced dataset in the diabetes category.

E. MORTALITY RISK PREDICTION (ICU)

Four studies focused on the mortality risk prediction category in ICU settings, i.e., qualitative analysis in pediatric intensive care units [64], overall mortality risk prediction [62], [65], and assessing the risk of extubation failure in ICU patients undergoing vitalisation [63]. All datasets used are private and include patient vitals, hospital records, and laboratory tests. These datasets are often imbalanced, with missing values and outliers that require preprocessing to ensure accurate model predictions.

F. PNEUMONIA, PREGNANCY, AND ENDOSCOPY

Three studies each focused on pregnancy [74]–[76] and endoscopy [66]–[68], while two studies addressed pneumonia [58], [78]. In pregnancy, studies predicted preterm births [75], extrauterine growth restriction [76], and anticipated cesarean delivery outcomes [74]. Endoscopy and pneumonia studies are primarily used and publicly available image-based, except for one private pneumonia disease dataset. Two of the endoscopy datasets are multiclass, and one is highly imbalanced. The pregnancy category dataset is mostly tabular, imbalanced, and contains missing values.

G. RARE APPLICATIONS

Less frequently studied application areas include skin lesion classification [69], skin vascular wound images [70], antimicrobial [73], [79], hepatitis liver disease [79], pulse wave classification [80], obesity [50], surgical [46], chronic [82] and acute disease [81]. In the context of antimicrobial aspects, early detection of drug resistance was studied in [72], and antimicrobial stewardship was examined in [73]. These studies utilised both tabular (see Table 8) and image datasets (see 9), with a mix of public and private sources, with issues of imbalance and missing data. Despite their rarity, these applications highlight the versatility of XAI methods in addressing diverse medical conditions.

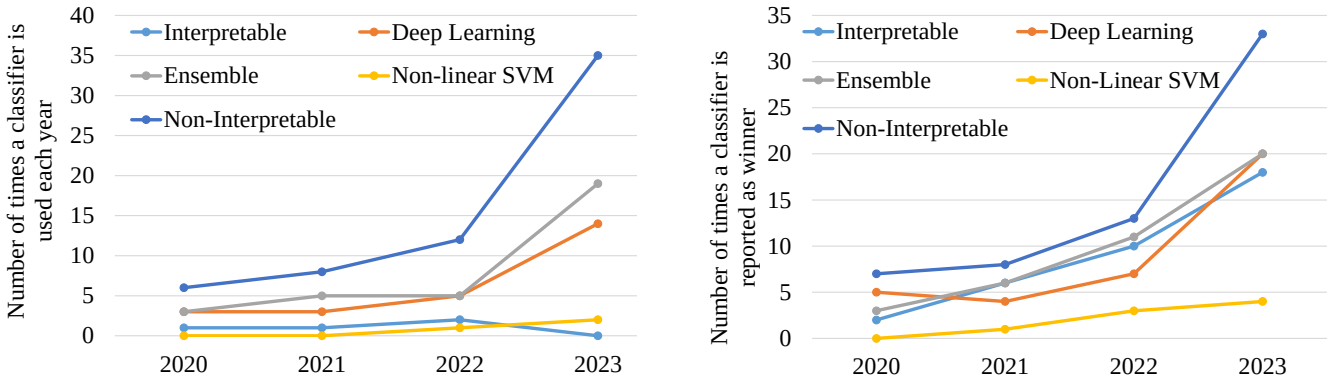
From the above discussion, tabular datasets are mostly imbalanced with missing values and noise, necessitating preprocessing and balancing for improved results. Common methods used for handling missing values when more than 30% is missing, include dropping data [20]–[22], [26], [30], [31], [33], [41], [50], [53], [54], [75], [77], [82], means, median or mode imputation [26], [30], [51]–[54], [60], [63], [76], [77], [84], KNN interpolation [28], [33], [74] and last observation carried forward [63], [65]. However, [52] noted that mean is sensitive to noise and results in wrong imputations. Other methods include iterative imputation [56], [61], baseline wander removal and power line interference removal [47] and forward imputation [27]. For data balancing, SMOTE [30], [36], [60], [62], [75] and its other variations such as SMOTE-ENN [48], [49], borderline SMOTE [57], SMOTE-Tomek [57] and SMOTE-NC [62] are the most common methods applied in the literature. Other common techniques applied are random subsampling [22], [75], under-sampling [75] and ADASYN [57], [74]. It is also observed that most of the datasets among tabular categories are private, particularly in the neurological category, where 12 out of 17 datasets are private. Similarly, all the datasets related to mortality risk prediction, anti-microbial, and Pulse wave classification are private. Likewise, more than 50% datasets in all the categories are private.

TABLE 7: Categorisation of the best-performing ML models and the frequency of each model and category being reported as the best-performing.

Category	Best Performing ML Models	References
Interpretable (4)	LR (3)	[39], [44], [77]
	Elastic Net (1)	[40]
Deep Learning (26)	CNN (15)	[23], [34], [45]–[47], [58], [66]–[71], [78], [81], [83]
	LSTM (3)	[80], [72], [65]
	Light_LSTM (1)	[63]
	BO-Tabnet (1)	[51]
	RNN (1)	[27]
	HAE-TabNet (1)	[49]
	MLP (2)	[20], [50]
	Mental RoBERTa (1)	[24]
	DNN (1)	[82]
Ensemble/ Stacking (32)	RF (13)	[26], [28], [32], [36], [38], [54], [55], [57], [62], [64], [74], [79], [84]
	XGB (6)	[25], [31], [33], [42], [48], [50]
	CatBoost (2)	[22], [29]
	LightGBM (1)	[87]
	GBDT (2)	[59], [73]
	GB (1)	[61]
	AdaBoost (3)	[21], [41], [52]
	Ensemble (RNN, GRU, BiLSTM) (1)	[43]
	Ensemble (LR, DT, MLP) (1)	[35]
	Ensemble (XGB, LGB, SVC, RF, CatBoost) (1)	[85]
	Ensemble (ConvXGB) (1)	[?]
	Stacking (RF, DT, KNN, LR, NB, SVM) (1)	[30]
	Stacking (AdaBoost, CatBoost, LGB, XGB) (1)	[60]
	Voting ensemble (1)	[75]
SVM (3)	Non-Linear SVM (3)	[32], [53], [76]

IV. PREDICTIVE MODELS

Machine learning models are broadly classified into interpretable and non-interpretable models. Interpretable models are simple and transparent, allowing humans to easily understand their predictions or decisions. Examples of interpretable models include Logistic Regression, Decision Trees, and Naive Bayes. On the other hand, non-interpretable models are complex and lack transparency in decision-making, thus requiring an explainer method for explaining decision-making. Examples of non-interpretable models include non-



(a) Models Applied

(b) Winner Models

FIGURE 4: Baseline ML models and winners' trends from Jan 2020 to Dec 2023 are shown; data for 2024 is not shown as it is an ongoing year.

linear SVM, ensembles, and deep learning.

Figures 4a and 4b illustrate the evolution of models used and winning models in literature from 2020 to 2023. 2024 data is not included as the year is ongoing. The trend indicates that interpretable and non-interpretable models are preferred as baseline models, but as time passes, non-interpretable models are increasingly observed as winners, with a significant spike in 2023 and a complete drop for interpretable to zero in 2023. Additionally, Table 7 lists the categories of models and their frequency count as best-performing, along with an individual model frequency count as winners. Ensembles/stacking emerge as the most prevalent category, followed by deep learning, highlighting non-interpretable models as the top performers overall between 2020 and 2024.

As shown in Table 7, a diverse range of ML models is evident, warranting further discussion to provide a concise overview. Convolutional Neural Networks are ideal for scenarios involving image data such as skin lesion classification, endoscopy, brain diseases, cancer, viruses, cardiovascular diseases, and electronic health record time-series data. LSTM is competitive on time-series and sequential data [65], [72].

Random Forest is a robust choice for structured tabular medical data, demonstrating high performance in predicting liver disease, diabetes, brain, cardiovascular, cancer, COVID-19, and mortality risk prediction time-series data. Ensembles are competitive due to their ability to aggregate predictions from multiple models, rendering them robust and highly effective across various medical domains. For example, an ensemble of deep learning techniques excels on image datasets [35], [43], while an ensemble of tree-based models performs best on tabular datasets [56]. Deep learning has outperformed XGBoost, Naive Bayes, Random Forest, and CatBoost on the tabular dataset [85].

Support Vector Machines (SVMs) have stood out as a potent model for medical predictions, being successful in gestational diabetes prediction [53], predicting the functional outcome of stroke survivors [32], and detecting prostate cancer on cancer images datasets [38].

Among interpretable models, logistic regression and Elastic NET are chosen as the baseline and best-performing models, which have shown noteworthy performance on tabular data in specific clinical contexts such as cancer [39], [44], [77] and Cytokines analysis [77].

TABLE 8: Overview of Tabular Datasets by Application Area, Availability and Types of Data

App Areas	Name and ref.	Features Description	Missing Values/ Outliers	Avail	Class imbalance	Dataset size and classification type	Sampling strategy	Strategy for missing values
Neurological	aMCI conversion to dementia [25]	Demographics and clinical characteristics	Not mentioned	Prv	Imbalanced	19X705 (Binary)	-	Domain knowledge to remove the unnecessary variables
	PPMI dataset [26]	Subject Characteristics, Bio-Samples, Medication History, Motor function, and Non-Motor functions	Yes	ARQ	Imbalanced	5X1059 (Multiclass)	SMOTENN	>30% were dropped. Forward and backward filling, median and the mode
	TRACK-TBI (EHR and Physiological) [27]	EHR: Vitals, lab measurements, GCS score components Physiological: Vital signs and intracranial data	EHR: 22%, Physiological: 8.8%	ARQ	N/A	EHR: 12x900 (Multi), Physio: 8x5816 (Binary)	-	EHR: GRU-D units, Physiological: Linear interpolation
	MIMIC III EHR [27]	Vital signs & lab measurements	10%	Pub	N/A	26X22988 (Multiclass)	-	Forward imputation
	ADNI [30]	Clinical dementia rating, functional activities questionnaire test, AD assessment scale, and Demographics	Yes	ARQ	Imbalanced	30X1363 (Multiclass)	SMOTE	[30] :Median, mean and mode. [28]: >30% dropped, and KNN imputation
	HCP [29]	Social and Emotional processing tasks	N/A	ARQ	N/A	246X30135 (Multiclass)		
	Cerebral Stroke Prediction- Imbalanced [20]	Demographic and medical history	Outliers and Missing	Pub	1.8%:98%	10x43,400 (Binary)	-	Dropped
	MDD-BD [31]	Sociodemographics, past history, vital signs, laboratory tests and chief complaints	Yes	ARQ	Imbalanced	93X16311 (Multiclass)	-	Dropped
	Ischemic stroke [22]	Age, Sex, History of cardiac & diabetes mellitus, Hypercholesterolemia presence, and Thrombolysis treatment	Yes	ARQ	28%:72%	7X514 (Binary)	Random sub-sampling	>5% Dropped
	Stroke Survivors [32]	Demographics, medical history, type of stroke, admission levels of systolic blood pressure, glucose, CRP and ESR	N/A	Prv	N/A	35X470 (Binary)		
	Ischemic stroke [84]	Demographics and medical history	Yes	Pub	4.9%:95.1%	10X5110 (Binary)	-	Mean
	Mental health survey 14 & 16 (Kaggle) [21]	Demographics, work environment and co-workers, family history, wellness	D1: Mismatched D2: >60% missing (16 var)	Pub	50%:50%	27X1259, 63X1433 (Binary)	-	Dropped
	Understanding Society [21]	General Population Sample of the UK Household Longitudinal Study	N/A	Prv	22%: 62%: 16%	330X11745 (Multiclass)		
	Chronic disease prediction [33]	Routine blood and biochemical test	Yes	Prv	30%: 53.5%: 16.5%	37X32448 (Multiclass)	-	>10% dropped and kNN interpolation

Cancer	Cervical cancer behavior risk (UCI) [35]	Eating and hygenic behaviour, perception, attitude, social support etc	No	Public	30%:70%	19X72 (Binary)		
	Lungs Cancer [41]	Features extracted from lung CT images	Yes	Prv	56.2%: 43.8%	286X2063 (Binary)	-	Columns and rows >50% missing values were dropped. Forward-fill and back-fill for remaining
	Survey Lung Cancer	Demographic, Smoking, Yellow fingers, Anxiety, Peer Pressure, Fatigue, Allergy, Wheezing, Shortness of Breath, Swallowing Difficulty, Chest Pain	N/A	Pub	Imbalanced	17x319 (Binary)	SMOTE	
	Multiple myeloma dataset [36]	Demographic information, personal and family history, various analysis results, medical examinations and diagnostic tests	N/A	Pub	Imbalanced	57X102 (Multiclass)	SMOTE Multiclass	
	Surveillance, Epidemiology, and End Results (SEER) [42]	Demographic, AJCC Stages, Grade, Chemotherapy, Surgical resection and Radiotherapy	N/A	ARQ	40%: 60%	11X1094 (Binary)		
	Breast cancer [21]	Characteristics of the cell nuclei present in the image	No	Pub (UCI)	63%:37%	30X569 (Binary)		
	Chronic disease prediction [33]	Routine blood and biochemical test	Yes	Prv	30%: 53.5%: 16.5%	37X32448 (Multiclass)	-	>10% were removed. KNN interpolation
Cardiovascular	CHF-CRF [48]	Demographics, medical history, physicals status, echocardiography, electrocardiography, and laboratory parameters	Yes	ARQ	Imbalanced	8X5188 (Binary)	SMOTE-ENN	
	Cardiac Arrest Survival [49]	Demographic, past history, insurance, place of cardiac arrest and other medical characteristics	Yes	Prv	87.2%: 12.8%	216X30179 (Binary)	SMOTE-ENN	Columns >50% null values were eliminated.
	China Physiological Signal Challenge 2018 [47]	Features from ECG recordings, age and sex	Noise	Pub	N/A	7X6,877 (Multiclass)	-	Baseline Wander Removal and Power Line Interference Removal
	ECG200 and Synthetic [46]	ECG200: Electrical activity recorded during a single heartbeat, Synthetic: ECG measurements	N/A	Pub	ECG200: 67%: 33%, Synthetic: 50%:50%	ECG200: 200, Synthetic: 1000 (Binary)		
	Cardiovascular Disease Dataset [50]	age, height, weight, gender, systolic BP, diastolic BP, cholesterol, glucose, smoking, alcohol, physical activity	Yes	ARQ	Balanced (48.8%: 51.2%)	11X70,000 (Binary)	-	Dropped
	Cardiotocography [21]	Cardiotocograph Features	No	Pub (UCI)	78%:14%: 8%	21X2126 (Multiclass)		
	Chronic disease prediction [33]	Routine blood and biochemical test	Yes	Prv	30%: 54%: 16%	37X32448 (Multiclass)	-	>10% were removed and kNN interpolation
Diabetes	GDM [53], [54]	Clinical data collected at the PEARS	Yes	Prv	Highly imbalanced	23X565 (Binary)	SMOTE	>30% dropped. Median, mean and mode.

	GDM [52]	Age, Ethnicity, Diabetes mellitus, BP (mmHg), Central armellini fat, Current gestational age, Pregnancies, First fasting glucose, BMI pregestational, Gestational age at birth, Type of delivery, Child's birth weight	Yes	Pub	13%: 87%	13x133	SMOTE, ADASYN, SMOTE-(ENN, Tomek, borderline)	Median
	ESDRPD [51]	Age, gender and other features collected from patients via questionnaires	No	Pub	38.5%: 61.5%	16X520 (Binary)		
	DR prevalence detection [55]	Demographic, Family history and clinical observations	N/A	Prv	N/A	10X172 (Binary)		
	Pima indians diabetes [51]	Glucose, BP, Skin Thickness, Pregnancies, insulin, BMI, Age, Diabetes pedigree function	Missing values and extreme outliers	Pub	65%:35%	8X767 (Binary)	-	Median
	Diabetic retinopathy [21]	Features extracted from the Messidor images	No	Pub	53%:47%	19X1151 (Binary)		
COVID-19	CHESS database [59]	demographics and risk factors on patients with a confirmed diagnosis of COVID-19	N/A	Prv	65%: 35%	19X13954 (Binary)		
	COVID-19 [56]	Age, BMI, sex, alcohol, cannabis, contacts count, COVID19 symptoms, smoking and different chronic diseases	13.8% missing	Pub	98.8%: 1.20%	59X1023426 (Binary)	Under sampling (1:3)	Iterative imputation techniques.
	COVID-19 dataset [60]	Demographics and clinical observations	Yes	Prv	23%: 77%	22X1169 (Binary)	Borderline-SMOTE	Mean, median and mode. IQR for handling outliers.
	COVID-19 dataset (Kaggle) [57]	Demographic parameters, nine grouped diseases, blood parameters, and vital signs of COVID-19-positive	Yes	Pub	26.6%: 73.4%	231X1925 (Binary)	ADASYN, SMOTE, SMOTE-(Tomek, Borderline, and ENN)	
	COVID-19 dataset [61]	Demographics, clinical signs, chronic illnesses and platelet-disrupting medications within the previous two weeks	>50% missing for 29 rows	ARQ	65.5%: 34.5%	48X87 (500 samples of synthetic data) (Binary)	-	Iterative imputation using chained equations Forest (Mice-Forest)
Anti-microbial	Antimicrobial multidrug resistance (AMR) [72]	Epidemiology, emergence, prevalence, and infectious diseases	Yes	ARQ	18%: 82%	23X3470 (Binary)	Under sampling and Balanced Cross-Entropy	
	Antibiotic prescriptions and susceptibilities [73]	Admission data, patient demographics (age and sex), prescription, and clinical records of culture tests	Yes	ARQ	N/A	51X5190 (Binary)		
Mortality risk (ICU)	MIMIC-III [62]	Patient's vital signs, hospital records, fluid information, laboratory tests, treatment orders, and free-text medical records	Yes	ARQ	98.5%: 1.5%	460X7874 (Binary)	SMOTE, SMOTE-NC	Proposed customised algorithm

	CHP Hospital Dataset [64]	Demographics, Hospitalisation data, Assigned diagnoses, Recorded locations, Ventilation, Physical assessment and Laboratory test results	Yes	ARQ	N/A	422 (Binary)		
	Danish National Patient Registry (DNPR) [65]	Demographics and diagnoses (daily obtained information, and data obtained with high sampling rate)	Yes	Prv	Imbalanced	44X15615 (Binary)	-	Last observation carried forward (LOCF)
	MIMIC-IV [63]	Demographic characteristics and clinical features	Yes	ARQ	70%:30%	89X8599 (Binary)	-	Last observation carried forward and Mean imputation for without any observation
Pregnancy	HosmartAI project Dataset [75]	Demographics, social and medical history, and obstetrics variables	Yes	Prv	34%:66%	32X375 (Binary)	Random under-sampling and SMOTE	>30% were dropped, and others with the most frequent or median.
	CHA Bundang Medical Center ICU [76]	Demographic data and the initial assessment results (vital signs, imaging findings, and laboratory tests)	Yes	ARQ	35%: 65%	26X124 (Binary)	-	Mean imputation
	PDHS'17-18 and '12-13 (C-section) [74]	Household information, contraceptive knowledge and practice, post-delivery, Children's health care, Nutrition and migration patterns	Yes	Pub	13.8%: 86.2%	875X15,409 (Binary)	ADASYN	KNN imputation
Hepatitis	Hepatitis Dataset (UCI) [79]	Demographics and clinical observations	48% instances missing	Pub	20.6%: 79.4%	19X155 (Binary)	SMOTE	
Surgical	SSI (Surgical) [46]	C-reactive protein	N/A	Prv	73.6%: 26.4%	883 (Binary)		
Chronic and Acute	NHANES (Chronic) [82]	Demographic, socioeconomic, Selected medical & laboratory tests, and self-reported data	Missing Values and outliers	Pub	66%: 34%	51X19225 (Multiclass)	-	Dropped missing values and outliers. Interquartile range for outlier elimination.
	CROSS-TRACKS (Acute) [81]	Laboratory parameters and vital signs	Yes	ARQ	Imbalanced	33X163050 (Multiclass)	Over-sampling	Standard carry-forward interpolation
Cytokines analysis	HIV-DED Dataset [77]	Cytokine-related characteristics and a binary feature indicate which eye is involved	10% for 2 variables and 49% for 4	Prv	N/A	126X42 (Binary)	-	>10 were Dropped. Imputed through the mean.
Obseity	Diabetes BRFSS 2015, Cardiovascular and Heart disease datasets [50]	Demographic and clinical	Outliers and duplicates	Pub	Diabetes: 85%: 15%, Cardio: 49%, 51%, HD: 91%, 9%	Diabetes: 21X253680, Cardio: 11X70000, HD: 279X400000 (Binary)		

TABLE 9: Overview of Image Datasets by Application Area, Availability and Types of Data

App Area	Name	Description	Availability	Dataset size	Multi/ Binary class	Class balance	Other Remarks	Ref.
COVID-19	NIH	CXR image	Public	112120 (1024 x 1024)	Multiclass	-	-	[58]
	SARS-COV-2 Ct-Scan	CT Scan	Public	2481 (244x244)	Binary	Balanced	Image resizing	
Endoscopy	Kvasir-capsule dataset	Endoscopy images	Public	47,238 images and 117 videos	Multiclass	Highly Imbalanced	Image resizing, normalisation and image augmentations with vertical and horizontal flips	[68]
	Kvasir	Endoscopy images (gastrointestinal tract)	Public	470,000 (720X579 to 1920X1070)	Multiclass	-	Data augmentation	[67]
	Red Lesion Endoscopy	Video capsule endoscopy images	Public	3295 (320x320)	Binary	34%:65%	Preprocessed but no discussion on preprocessing	[66]
Cancer	TCGA-GBM	MR, CT, DX	Public	481,158	Binary	-	-	[40]
	Brain tumor	T1-weighted MRI	Public	3064 (512×512)	Binary	-	Data augmentation	[39]
	Breast cancer (Valparaíso, Chile)	Histological samples	ARQ	1,000×750	Multiclass	-	-	[44]
	D1: colorectal cancer and D2: osteosarcoma dataset	MR Images	Private	D1: 165 (567 × 430 to 775 × 522) D2:1144	D1: Binary, D2:Multiclass	Imbalanced	-	[43]
	Cancer Imaging Archive	MRI and US image data	Public	611119	Binary	-	-	[38]
	Thyroid dataset	Ultrasound images	Private	19341	Binary	Highly Imbalanced	-	[45]
Pneumonia	Taichung Veterans General Hospital Dataset, Taiwan, NIH and VinDr	CXR image	D1:Private, D2:Public, D3: Public	D1:2301, D2:112120 (1024 x 1024), D3:18000	Multiclass	-	Data augmentation	[78]
	NIH	CXR image	Public	112120 (1024 x 1024)	Multiclass	-	-	[58]
Pulse Wave Classification	China Medical University Hospital Dataset	Raw Pulse Wave	Private	-	-	Imbalanced	Oversampling	[80]
Skin	ISIC 2019 dataset	Dermoscopic images	Public	25331 Images	Multiclass	Highly Imbalanced	Data augmentation (Both)	[69], [71]
	Vascular wound image registry	Vascular wound images	ARQ	2957	Multiclass	Imbalanced	Oversampling and augmentation techniques	[70]
Neurological	Felipe Kitamura's CT dataset	CT Images	Public	200 (512X512)	Binary	Balanced	-	[23]
	ADNI	3D MRI	ARQ	1,692	Binary	-	-	[34]

V. XAI METHODS

Explainable methods can be defined based on five aspects [88], *stage*, *applicability*, *scope*, *form* and *type*. The stage of explanation can be post hoc or ante-hoc. Post hoc methods are applied after the ML model is built (e.g., LIME). In contrast, self-explaining or ante-hoc methods are inherently explainable by design (e.g., AGRAD). In terms of applicability, post hoc methods can be model-agnostic or model-specific. Model-agnostic methods can explain and be applied to any ML model (e.g., LORE), and model-specific models can only be applied to specific ML models (e.g., LRP for deep learning). The method can have either a global (explaining the general behavior of the ML model) or local scope (explaining single instances or predictions). Furthermore, the explainer method can present explanations in various forms, such as rule-based interpretations (e.g., Anchors) or visual representations highlighting key aspects (e.g., GradCAM). The output of an explainer can belong to different types, like plots (e.g., PDP), graphs (e.g., Qlattice), feature importance (e.g., Saliency Map), or contrastive comparisons (e.g., SHAP).

Table 10 presents a comprehensive overview of the XAI methods utilised in the literature, outlining various aspects of each explainer method and the categories of ML models to which these methods were applied, along with references. It can be seen that the most widely used explainer method is SHAP, which is applied fairly across different categories of ML models. Following SHAP, LIME emerges as the second most popular choice, also employed in a similar manner, and both are model-agnostic methods. Among model-specific methods, GradCAM stands out as a preferred option for deep learning models. In terms of *stage*, the majority of methods are post hoc, with the exception of two methods, namely AGRAD and TabNet, both utilised in conjunction with deep learning and are attention-based methods. Concerning the *scope*, the emphasis is primarily on local methods, highlighting the significance of explaining individual instances or patient records as opposed to the broader behavior of a condition, class, or disease, although global scope methods are also utilised. Visualisation is the preferred output format among the selected methods, providing a straightforward means of explaining decisions, followed by a rule-based approach that explains reasoning through feature constraints. Finally, the output type is largely determined by the importance of features in influencing a decision, typically involving a comparison of each feature's impact on an automated decision (such as identifying correlated or influential features in a specific decision). Additionally, there are occasional instances of graphs and plots that explain the decision-making process.

VI. STUDIES UTILISING MULTIPLE XAI METHODS:

Twenty-five studies have utilised various explainable methods to explain prediction outcomes, with only ten comparing XAI methods. Six of these ten studies focused on image data, while four analysed tabular data. Deep learning techniques were used in all studies for image data, applying between three to six XAI methods. Heatmap was the dominant comparison strategy, followed by user studies, saliency maps, and explainability scores. GradCAM and its variant were the most favoured XAI methods, followed by LIME, SHAP, CIU, and ABELE, the once preferred method in different studies. For tabular data, ensemble methods were used across all studies, with one incorporating deep learning and non-linear SVM in addition to ensembles. All studies utilised two to three XAI methods, highlighting AdaWhip and BoCSoR as preferred methods, while others viewed XAI methods as complementary and supportive of their respective goals.

Table 11 presents a comprehensive overview of studies that utilised multiple XAI methods and contrasted them in their research. For image data,

- In [58], LIME achieved better quantitative scores than other XAI methods. However, radiologists preferred the Ensemble XAI over other methods for its localisation, effectiveness, and trust based on evaluation of heatmaps and computational complexity.
- In [67], GradCAM++ was identified as the top-performing method for endoscopic analysis based on heatmap evaluation. Similarly, in [68], GradCAM was found to be superior for visual explanations during backpropagation, while SHAP and LIME were also acknowledged for their efficacy in feature-based explanations.
- In [66], LIME, SHAP, and CIU were evaluated based on human comprehension, satisfaction scores, computational complexity, and overall understanding. CIU was highlighted as the superior method compared to LIME and SHAP.
- In [70], SHAP was favored over LIME and GradCAM for explaining decisions based on an *explainability score* that measured the model's focus on the wound area compared to the rest of the image.
- In [71], ABELE was more effective than local explainers LIME and LORE in skin lesion detection due to its superior saliency maps despite higher computational complexity.

For tabular data,

- In [21], Anchors, AdaWhip, and LORE were compared based on mean coverage and mean precision, with AdaWhip performing better than the other two.
- In [36], SHAP, LIME, and permutation feature importance (PFI) were evaluated based on feature importance values and computational complexity. PFI emerged as the most efficient method for general understanding, while LIME and SHAP proved valuable for individual instances, with consensus among methods underscoring the significance of distinctive features for each cancer patient.
- In [29], a novel XAI method, BoCSoR, was compared with SHAP based on average correlation values for social and emotional tasks, with BoCSoR outperforming

TABLE 10: Methods for Explainability. Abbreviations by column *Applicability*=App., Agnoistic=agn, Specific=spe; *Stage*=St., post hoc=p, ante hoc=a; *Scope*=Sc., local=l, global=l; visual=vis; feature importance=fi, contrastive=con, plot=plt.

XAI Method	App.	St.	Sc.	Form	Type	Model Cat.	References
SHAP (43)	agn	p	l/g	vis	fi/con	Interpretable (6)	[26], [29], [40], [44], [76], [80]
						Ensemble (29)	[22], [25], [26], [28]–[33], [35], [36], [42], [48], [50], [53], [55]–[57], [59]–[62], [64], [73]–[75], [79], [80], [84]
						Deep Learning (15)	[20], [29], [32], [50], [51], [58], [63], [65], [66], [68], [70], [72], [78], [80], [82]
						Non-Lin. SVM (1)	[76]
LIME (24)	agn	p	l	vis	fi/con	Interpretable (2)	[26], [77]
						Ensemble (12)	[35], [36], [41]–[43], [54], [57], [60], [61], [74], [79], [84]
						Deep Learning (10)	[24], [38], [49], [51], [58], [66], [68]–[71]
GradCAM (8)	spe	p	l	vis	fi	Deep Learning (8)	[23], [45], [47], [58], [67], [68], [70], [83]
Feature Importance (5)	spe	p	l/g	vis	fi	Interpretable (2)	[22], [77]
						Ensemble (4)	[22], [25], [57], [74]
ELI5 (2)	agn	p	l/g	vis	fi/con	Ensemble (2)	[57], [60]
Qlattice (2)	agn	p	g	rule	graph	Ensemble (2)	[57], [60]
Anchors (2)	agn	p	l	rule	-	Ensemble (2)	[21], [57]
GradCAM++ (3)	spe	p	l	vis	fi	Deep Learning (3)	[58], [67], [68]
Saliency MAP (1)	agn	p	l	vis	fi	Deep Learning (1)	[58]
LRP (2)	spe	p	l	vis	fi	Deep Learning (2)	[22], [81]
Graph (2)	agn	p	g	vis	graph	Interpretable (1)	[39]
						Ensemble (1)	[50]
						Deep Learning (1)	[50]
LORE (2)	agn	p	l	rule	con	Ensemble (1)	[21]
						Deep Learning (1)	[71]
PDP (2)	agn	p	g	vis	plt	Ensemble (2)	[25], [79]
PFI (2)	agn	p	g	vis	fi	Interpretable (1)	[77]
						Ensemble (1)	[36]
ABELE (1)	agn	p	l	vis	fi/con	Deep Learning (1)	[71]
BoCSoR (1)	agn	p	g	vis	fi/con	Interpretable (1)	[29]
						Ensemble (1)	[29]
						Deep Learning (1)	[29]
CIU (1)	agn	p	l	vis	fi/con	Deep Learning (1)	[66]
LayerCAM (1)	spe	p	l	vis	fi	Deep Learning (1)	[68]
CAM (1)	spe	p	l	vis	fi	Deep Learning (1)	[46]
Guided GradCAM (1)	spe	p	l	vis	fi	Deep Learning (1)	[34]
AGRAD (1)	spe	a	l	vis	fi/con	Deep Learning (1)	[24]
ICE (1)	agn	p	l	vis	plt	Ensemble (1)	[25]
Ada-WHIPS (1)	spe	p	l	rule	-	Ensemble	[21]
WindowSHAP (1)	agn	p	l	vis	fi/con	Deep Learning (1)	[27]
Ensemble XAI (1)	spe	p	l	vis	fi	Deep Learning (1)	[58]
Break Down (1)	agn	p	l	vis	fi/con	Ensemble (1)	[31]
TabNet (1)	spe	a	l/g	vis	fi	Deep Learning (1)	[51]

SHAP in feature correlation values.

- In [61], LIME and SHAP were compared based on feature importance values for COVID-19 severity in patients. The common key features identified by both methods were deemed robust indicators for each patient's record.

Fifteen additional studies have employed multiple XAI methods without comparing their performance but underscored their usefulness. Among these, 14 studies focused on tabular data, with only one involving text data. These studies and results are outlined in Table 12 and the following shows key characteristics of these studies:

- Studies [25], [31], [57], [77], [79], [84] confirmed the usefulness of XAI methods by confirming important identified features through previous research.
- Some studies have shown that XAI methods can be beneficial in pinpointing crucial features related to a medical condition for which a definitive solution is yet to be discovered, as seen during the COVID-19 pandemic [60]. They can also be instrumental in unexplored research domains like early detection of Parkinson's disease [26], assessing cancer risk [35], cancer survival [42], supporting healthcare professionals and patients [50], and addressing mental health conditions [24], [31].
- In [74], XAI methods were proposed as an early explainable predictive approach to aid in implementing new policies to reduce unnecessary C-Section deliveries.
- In [51] demonstrated the effectiveness of collective inference using ante-hoc and posthoc XAI methods, and in [22], various XAI methods were employed to explain distinct ML models.

VII. RESEARCH GAPS AND RECOMMENDATIONS

The survey provides a comprehensive guide for medical practitioners and researchers, providing an in-depth review of predictive models, XAI methods, current trends in their adaptability, and datasets used. Based on our review and analysis of the studies, the following gaps have been identified.

1) *Limited availability public tabular datasets* - Over 55% of the datasets used in the reviewed studies are private, particularly in critical areas such as neurological disorders (only 3 out of 14 datasets are public), ICU mortality risk (all private), and antimicrobial (all private) category. This lack of public datasets hampers the development of replicable models and limits comparative studies.
2) *Lack of effective data treatment methods* - Many tabular datasets contain missing values, noise, and outliers. Common strategies like dropping data with significant missing values are not always effective, especially when dataset sizes are already limited.
3) *Lack of feature selection and engineering* - Feature engineering is crucial for improving classifiers' predictive performance. Although it is underutilised in CDSS, it is widely utilised in other ML applications.
4) *Data imbalance issue* - The data is not evenly distributed among the different classes in most of the dataset, making it challenging to train and evaluate models. This can lead to models being biased towards the majority class, resulting in poor performance for the minority class. Some studies suggest that data balancing techniques improve predictive performance, while others find little impact on predictive performance but note a widening of output probabilities demonstrating confidence of classification [81].
5) *The lack of reliable and automated solutions* - There is still a lack of reliable solutions that provide convincing explanations to medical experts in their decision-making. This is due to the ethical and regulatory concerns about the use of AI in healthcare, including biases, lack of transparency, privacy concerns, and safety and liability issues. Ensembles and deep learning-based approaches have proven helpful for building highly performing models; however, they lack interpretability and transparency.
6) *Diverse Application Areas but Uneven Coverage* - While some application areas, like neurological conditions and cancer, are well-studied, others, such as antimicrobial resistance, hepatitis, and obesity, receive less attention.
7) *Limited Use of Multiple XAI Methods* - Currently, there is a shortage of solutions that can provide convincing explanations to medical experts in their decision-making process, which is crucial for building trust and accountability among clinicians and patients. While some studies compare the effectiveness of different XAI methods, this comparison is mostly focused on image data. Limited formal comparisons exist for XAI metrics in tabular data. Additionally, many studies rely on single XAI methods, potentially missing out on a comprehensive understanding of interpretability.
8) *XAI Evaluation Metrics and Standards for XAI* - There is a lack of standardized evaluation metrics for assessing the effectiveness of XAI methods.

To address the discussed gaps, the following are the recommendations to follow:

1) *Openness of datasets* - Prioritise providing open access to high-quality medical datasets, anonymising patient data, obtaining patient consent for data privacy, setting clear publishing guidelines, enforcing data governance practices, and increasing transparency and accessibility for XAI-CDSS research.
2) *Data treatment* - Develop and adopt more sophisticated, medical-domain-informed data preprocessing techniques to effectively handle missing values, noisy data, and outliers using advanced statistical and machine learning-based approaches.
3) *Adopt feature selection and engineering approaches* -

TABLE 11: Summary of studies that used multiple XAI methods and compared them.

Ref	Data Type	Model Cat.	XAI Methods Used	Comparison Methodology	Observation	Preferred XAI Method
[58]	Image	Deep Learning	(6) SHAP, LIME, GradCAM, GradCAM++, Saliency MAP, Ensemble XAI	Heatmap along with absence impact, localisation and trust	Developed an ensemble of SHAP and Grad-CAM++. While showing comparable quantitative results to other methods, the qualitative evaluation by radiologists indicated that the XAI ensemble is more effective in localisation (precision: 0.52, recall: 0.57, F1: 0.50, IOU: 0.36) and trusted method by the panel of radiologists (mean vote: 70.2%).	LIME (quantitative), Ensemble XAI (qualitative)
[67]	Image	Deep Learning	(5) GradCAM, GradCAM++, LayerCAM, Hires-CAM, XGRAD-CAM	Heatmap	All methods yielded comparable results, with Grad-CAM++ showing higher heat maps for dyed-p and polyp classes. Because dyed-p contained more fixtures in the image, it produced a more precise heat map compared to the polyp class.	GradCAM++
[68]	Image	Deep Learning	(5) SHAP, LIME, GradCAM, Grad-CAM++, Layer-CAM	Heatmap	Propagation-based models extract better visual explanations from neural networks. GradCAM is the best method due to its reliance on back-propagation gradients. However, it struggles with the localisation of objects with multiple occurrences of the same class, impacting results in detail-dense applications such as bacterial microscopic images.	GradCAM
[66]	Image	Deep Learning	(3) SHAP, LIME, CIU	User (non-medical) study	CIU is computationally at least 1.15 times more efficient than the other two methods and also satisfies more mean (at least 1.16 times more) and median users (at least 1.16 times more) with its explanation compared to the other two methods.	CIU
[70]	Image	Deep Learning	(3) SHAP, LIME, GradCAM	Explainability score	The explainability score is calculated by assessing the model's focus on the wound area versus the rest of the image during its prediction. This is done using two segmentation models: level 1 (wound, periwound perimeter, wound perimeter, and background) and level 2 (18 different wound segmentation classes). SHAP achieved mean explainability scores of 0.61 for wound classification, 0.68 for wound measurement, and 0.72 for wound segmentation.	SHAP
[71]	Image	Deep Learning	(3) LIME, ABELE, LORE	Saliency MAP and user study	ABELE excels in generating saliency maps. Key pixel deletion has a bigger impact on ABELE, indicating its regions hold more importance (lime: 0.736, lore: 0.711, ABELE: 0.461 mean deletion AUC scores). When adding pixels, AUC scores improve, emphasising ABELE's ability to highlight critical decision-making areas (lime: 0.417, lore: 0.471, ABELE: 0.748 mean insertion AUC scores). The user study also confirmed that the ABELE explanation was effective for both experts and non-experts. Both groups showed improved accuracy in classifying instances when presented with ABELE explanations, with experts showing an average increase of 9% and non-experts with an average gain of 9.5%.	ABELE
[21]	Tabular	Ensemble	(3) Ada-WHIPS, Anchors, LORE	Mean coverage and mean precision	Anchors has low mean coverage, and Ada-WHIPS and LORE are comparably performant. Furthermore, LORE has low mean precision in most datasets, and Ada-WHIPS and Anchors are comparable.	AdaWhip
[36]	Tabular	Ensemble, Deep Learning, Non-Lin. SVM	(3) SHAP, LIME, PFI	Feature-based approach	PFI is important for general trends, but LIME and SHAP are critical for instance-based explanations. LIME is less computationally expensive than SHAP, and the two methods can complement each other by agreeing on important features.	Methods complement
[29]	Tabular	Ensemble	(2) SHAP, BoSCaR	Avg feature importance and correlation between feature importance	BoCSoR is more reliable in identifying the most important features for classification in physiological data, less sensitive to feature correlation, and less computationally expensive.	BoCSoR
[61]	Tabular	Ensemble	(2) SHAP, LIME	Feature importance	The SHAP and LIME methods were employed to assess the experimental results on the link between cytokine storm and COVID-19 severity in patients, as well as the influence of various cytokines on severity. The shared key features from SHAP and LIME were specifically used as robust indicators for local explanations.	Methods complement

TABLE 12: Summary of studies that used multiple XAI methods but did not compare them.

Ref	Data Type	Model Cat.	XAI Methods Used	Observation
[57]	Tabular	Ensemble	(6) SHAP, LIME, FI, ELI5, Qlattice, Anchor	The XAI methods have been validated by identifying the important features, which were confirmed by previous research. The research discovered that anomalies in features such as respiratory rate, blood pressure, body temperature, calcium, and lactate levels positively contribute to patient severity.
[60]	Tabular	Ensemble	(4) SHAP, LIME, ELI5, Qlattice	XAI methods were used to find important markers for screening COVID-19 patients. The important markers found were albumin, ALT, basophil, and TWBC. The authors said that XAI methods can help healthcare professionals in situations like COVID-19 when the best solution is unknown and be helpful in the first screening of coronavirus patients.
[25]	Tabular	Ensemble	(4) SHAP, ICE, FI, PDP	FI and PDP helped scrutinize the most important global features concerning cognitive impairment, with ICE and SHAP allowing the interpretation of specific patient data. Furthermore, using graphs, patients can better understand the neuropsychological factors at risk, which is a step towards precision medicine. Key discoveries from the explanations include RCFT delayed recall, CDR-SOB, age, K-MMSE, COWAT-animal, education, SVLT delayed recall, RCFT copy time, and APOE genotype - all of which resonate with previous research.
[79]	Tabular	Ensemble	(3) SHAP, LIME, PDP	PDP and SHAP (via mean SHAP values) were used as global explainers, and LIME as a local explainer method. These explainer methods found that ascites, spiders, bilirubin, albumin, malaise, varices, and the SpleenPalpable feature had more impact than the others, which is in line with prior knowledge from hepatobiliary physicians, confirming the effectiveness of these XAI methods.
[77]	Tabular	Interpretable, Ensemble	(3) LIME, FI, PFI	In a stacked-based clustering approach, permutation importance was used to evaluate the significance of each clustering method in grouping a set of features to identify distinct patterns for accurate classification. Feature importance served as the global method, while LIME was employed as the local explainer. The features identified through explainers confirmed the results obtained and supported previous studies, particularly highlighting the significance of cytokines GRO, EGF, and IP-10 and their association with DED disease and seropositivity.
[74]	Tabular	Ensemble	(3) LIME, SHAP, FI	The SHAP method was used to assess overall feature importance globally, but it was not sufficient for understanding each individual patient. As a result, LIME explanations were demonstrated to provide a more detailed understanding. The feature "had_previous_c_section" was identified as one of the most important global features. LIME was used to identify contributing features, including "suffered_domestic_violence." XAI methods were argued to provide an early explainable predictive approach to assist in implementing new policies to reduce unnecessary C-Section deliveries.
[51]	Tabular	Deep Learning	(3) SHAP, LIME, TabNet	Uses SHAP and TabNet as global explainers and LIME as a local explainer. The collective inference suggests that insulin and polyuria are significant features associated with diabetes risk. TabNet, an ante-hoc method, also demonstrates high accuracies on various datasets.
[22]	Tabular	Ensemble, Interpretable, Deep Learning	(3) SHAP, FI, LRP	Took a feature-based approach and used Shapley values for ensembles, model coefficients for logistic regression and deep Taylor decomposition for deep learning to explain respective models.
[42]	Tabular	Ensemble	(2) SHAP, LIME	Globally, SHAP showed that age, T-stage, ethnicity, M-stage, marital status, and grade were key factors for NPC patient survival. Both LIME and SHAP methods demonstrated how each feature impacted individual predictions, aligning with the globally identified important features.
[84]	Tabular	Ensemble	(2) SHAP, LIME	SHAP was used as a global explainer and identified Age, Average Glucose Level, Work Type, Residence Type, Gender, and Ever Married as important features. This generally aligns with the views of experts in hepato-biliary and previous research. Lime was used to reveal local features for individual patients, highlighting instances where Age and Work Type were occasionally ranked higher than other features for stroke patients, echoing observations in globally important features.
[26]	Tabular	Ensemble	(2) SHAP, LIME	Used SHAP for global and local explanations and LIME for local explanations to detect Parkinson's disease early. The research compared the outputs of SHAP and LIME for RF and LightGBM models. XAI methods highlighted the NP3BRADY feature as the most important, while local explainers identified the MESEADLG feature as the best feature for both LGBM and RF models.
[35]	Tabular	Ensemble	(2) SHAP, LIME	LIME and SHAP were used as local explainers. It was suggested that providing explanations for decision-making to medical professionals in scenarios like predicting cancer risk serves the purpose, even if methods share some features and differ in others. Severity perception was most indicative of cervical cancer in two patient cases, confirming a general understanding.
[50]	Tabular	Ensemble, Deep Learning	(2) SHAP, Graph	The study focused on examining the links between obesity, diabetes, cardiovascular issues, and heart disease. It utilised SHAP to provide insights globally and locally. Local findings were showcased through three case studies featuring patients with positive test results for multiple diseases, negative results, and varied predictions for comorbidities. A multi-node graph was used to aid healthcare professionals and patients in understanding the progression of these conditions, with ICD-10 codes playing a key role in interpretation. The approach assists clinicians in forecasting pathologies associated with obesity and transitioning to long-term prevention and treatment plans.
[31]	Tabular	Ensemble	(2) SHAP, Break Down	Uses SHAP for global and Break Down as local explainers. Overall, explainers offer four insights. 1- clinical significance is highlighted through top features. 2- the higher values of features, such as myocardial enzyme spectrum markers and diabetes-associated markers, lead to BD, while lower values contribute to MDD. 3- potential for the new discovery. 4- offers general recommendations.
[24]	Text	Deep Learning	(2) LIME, AGRAD	Uses LIME and model-specific attention-based AGRAD explanation methods to identify the word categories relied on by models when making predictions. Explainers reveal that multi-task fusion models learned significant correlations between mental health conditions, emotions, and personality traits. Specifically, AGRAD analysis highlights the models' preference for words in specific LIWC categories. The research suggests that interpreting models can help detect a range of mental health conditions, such as ADHD, anxiety, bipolar disorder, and depression.

Utilise feature selection and engineering techniques to identify important characteristics in tabular datasets. Consider domain expertise to highlight important features and discard less significant ones.

4) *Data Imbalance* - Explore various data balancing techniques on different datasets to determine the most effective approach for addressing data imbalance. Investigate counterfactual-based data augmentation methods [89] and other advanced and novel techniques informed of medical constraints (when available).
5) *Explainability Centric Model Design* - Creating new ML models with built-in explainability features. This includes developing deep learning architectures or ensemble methods that prioritise transparency without significantly compromising performance. Additionally, advancements in posthoc explanation methods should be made to ensure higher *fidelity*.
6) *Focus the less investigated Application Areas* - Research should be encouraged in less-studied application areas to ensure a more balanced coverage and explore the potential of XAI in a broader range of healthcare applications. A way to encourage this is by developing and curating public datasets related to less-studied application areas (such as antimicrobial resistance, hepatitis, and obesity). Publicly available datasets can facilitate research by providing standardized data for developing and testing XAI methods.
7) *Promote the Use of Comprehensive XAI Evaluations* - There is a need for comprehensive and detailed criteria to evaluate and compare existing XAI methods. This will help identify strengths, weaknesses, and areas for improvement, guiding future development to ensure accuracy, reliability, and ease of use for medical practitioners. Researchers should utilise multiple XAI methods in their studies and conduct comprehensive evaluations considering different interpretability aspects, including comparative analyses. Such studies can reveal the strengths and weaknesses of different XAI methods, enhancing our understanding of their effectiveness. A standardised evaluation criterion would enable fair and objective comparison of different XAI methods, leading to a better understanding of the state-of-the-art in this field and helping to identify the best methods to use in practical applications. Further research is required to develop reliable, automated solutions that offer compelling explanations.
8) *Standardize XAI Evaluation Metrics* - More research is needed to develop and disseminate standardized evaluation metrics for XAI in CDSS. These metrics can form the basis for being widely accepted and used by the research community to facilitate the comparison of results across different studies.

In conclusion, the synthesis of datasets, winning models, and explainer methods form a rich tapestry in the evolution of XAI for CDSSs. Beyond the technical prowess, it emphasises ethical underpinnings and a commitment to transparency, laying the foundation for a responsible and collaborative future in healthcare AI.

ACKNOWLEDGMENTS

This publication has emanated from research conducted with the support of the Irish Research Council under award nos. GOIPG/2022/660 & GOIPG/2021/1354 and grant number 18/CRT/6183 from SFI Centre for Research Training in Machine Learning (ML-Labs) at Technological University Dublin with the financial support of Science Foundation Ireland under grant no. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Technological University Dublin.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest associated with this paper.

REFERENCES

[1] A. Wasylewicz and A. Scheepers-Hoeks, "Clinical decision support systems," Fundamentals of clinical data science, pp. 153–169, 2019.
[2] A. Alanazi, F. Al Rabiah, H. Gadi, M. Househ, and B. Al Dosari, "Factors influencing pharmacists' intentions to use pharmacy information systems," Informatics in Medicine Unlocked, vol. 11, pp. 1–8, 2018.
[3] M. R. Kronenfeld, R. C. Bay, and W. Coombs, "Survey of user preferences from a comparative trial of uptodate and clinicalkey," Journal of the Medical Library Association: JMLA, vol. 101, no. 2, p. 151, 2013.
[4] P. Papadopoulos, M. Soflano, Y. Chaudy, W. Adejo, and T. M. Connolly, "A systematic review of technologies and standards used in the development of rule-based clinical decision support systems," Health and Technology, vol. 12, no. 4, pp. 713–727, 2022.
[5] G. Feder, M. Eccles, R. Grol, C. Griffiths, and J. Grimshaw, "Using clinical guidelines," Bmj, vol. 318, no. 7185, pp. 728–730, 1999.
[6] F. Magrabi, E. Ammenwerth, H. Hyppönen, N. de Keizer, P. Nykänen, M. Rigby, P. Scott, J. Talmon, and A. Georgiou, "Improving evaluation to address the unintended consequences of health information technology," Yearbook of medical informatics, vol. 25, no. 01, pp. 61–69, 2016.
[7] A. M. 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, no. 11, p. 5088, 2021.
[8] Y. Du, C. McNestry, L. Wei, A. M. Antoniadi, F. M. McAuliffe, and C. Mooney, "Machine learning-based clinical decision support systems for pregnancy care: a systematic review," International Journal of Medical Informatics, p. 105040, 2023.
[9] B. Vasey, S. Ursprung, B. Beddoe, E. H. Taylor, N. Marlow, N. Bilbro, P. Watkinson, and P. McCulloch, "Association of clinician diagnostic performance with machine learning–based decision support systems: a systematic review," JAMA network open, vol. 4, no. 3, pp. e211 276–e211 276, 2021.
[10] L. Wang, Z. Zhang, D. Wang, W. Cao, X. Zhou, P. Zhang, J. Liu, X. Fan, and F. Tian, "Human-centered design and evaluation of ai-empowered clinical decision support systems: a systematic review," Frontiers in Computer Science, vol. 5, p. 1187299, 2023.
[11] S. Moazemi, S. Vahdati, J. Li, S. Kalkhoff, L. J. Castano, B. Dewitz, R. Bibo, P. Sabouniaghdam, M. S. Tootooni, R. A. Bundschuh et al., "Artificial intelligence for clinical decision support for monitoring patients in cardiovascular icus: A systematic review," Frontiers in Medicine, vol. 10, p. 1109411, 2023.
[12] Q. Xu, W. Xie, B. Liao, C. Hu, L. Qin, Z. Yang, H. Xiong, Y. Lyu, Y. Zhou, A. Luo et al., "Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review," Journal of Healthcare Engineering, vol. 2023, 2023.
[13] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," Computer Methods and Programs in Biomedicine, vol. 226, p. 107161, 2022.

[14] S. Ali, F. Akhlaq, A. S. Imran, Z. Kastrati, S. M. Daudpota, and M. Moosa, "The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review," Computers in Biology and Medicine, p. 107555, 2023.
[15] N. Prentzas, A. Kakas, and C. S. Pattichis, "Explainable ai applications in the medical domain: a systematic review," arXiv preprint arXiv:2308.05411, 2023.
[16] S. S. Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos, and H.-W. Liang, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," Informatics in Medicine Unlocked, p. 101286, 2023.
[17] J. Allgaier, L. Mulansky, R. L. Draelos, and R. Pryss, "How does the model make predictions? a systematic literature review on the explainability power of machine learning in healthcare," Artificial Intelligence in Medicine, vol. 143, p. 102616, 2023.
[18] S. Keele et al., "Guidelines for performing systematic literature reviews in software engineering," 2007.
[19] D. Moher, L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L. A. Stewart, and P.-P. Group, "Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement," Systematic reviews, vol. 4, pp. 1–9, 2015.
[20] C. Kokkotis, G. Giarmatzis, E. Giannakou, S. Moustakidis, T. Tsatalas, D. Tsiptsios, K. Vadikolias, and N. Aggelousis, "An explainable machine learning pipeline for stroke prediction on imbalanced data," Diagnostics, vol. 12, no. 10, p. 2392, 2022.
[21] J. Hatwell, M. M. Gaber, and R. M. Atif Azad, "Ada-whips: explaining adaboost classification with applications in the health sciences," BMC Medical Informatics and Decision Making, vol. 20, no. 1, pp. 1–25, 2020.
[22] E. Zihni, V. I. Madai, M. Livne, I. Galinovic, A. A. Khalil, J. B. Fiebach, and D. Frey, "Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome," Plos one, vol. 15, no. 4, p. e0231166, 2020.
[23] K. H. Kim, H.-W. Koo, B.-J. Lee, S.-W. Yoon, and M.-J. Sohn, "Cerebral hemorrhage detection and localization with medical imaging for cerebrovascular disease diagnosis and treatment using explainable deep learning," Journal of the Korean Physical Society, vol. 79, no. 3, pp. 321–327, 2021.
[24] E. Kerz, S. Zanwar, Y. Qiao, and D. Wiechmann, "Toward explainable ai (xai) for mental health detection based on language behavior," Frontiers in psychiatry, vol. 14, 2023.
[25] M. Y. Chun, C. J. Park, J. Kim, J. H. Jeong, H. Jang, K. Kim, and S. W. Seo, "Prediction of conversion to dementia using interpretable machine learning in patients with amnestic mild cognitive impairment," Frontiers in Aging Neuroscience, vol. 14, p. 898940, 2022.
[26] M. Junaid, S. Ali, F. Eid, S. El-Sappagh, and T. Abuhmed, "Explainable machine learning models based on multimodal time-series data for the early detection of parkinson's disease," Computer Methods and Programs in Biomedicine, vol. 234, p. 107495, 2023.
[27] A. Nayebi, S. Tipirneni, C. K. Reddy, B. Foreman, and V. Subbian, "Windowshap: An efficient framework for explaining time-series classifiers based on shapley values," Journal of Biomedical Informatics, vol. 144, p. 104438, 2023.
[28] S. El-Sappagh, J. M. Alonso, S. R. Islam, A. M. Sultan, and K. S. Kwak, "A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease," Scientific reports, vol. 11, no. 1, p. 2660, 2021.
[29] A. L. Alfeo, A. G. Zippo, V. Catrambone, M. G. Cimino, N. Toschi, and G. Valenza, "From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks," Computer Methods and Programs in Biomedicine, vol. 236, p. 107550, 2023.
[30] A. Almohimeed, R. M. Saad, S. Mostafa, N. El-Rashidy, S. Farag, A. Gaballah, M. Abd Elaziz, S. El-Sappagh, and H. Saleh, "Explainable artificial intelligence of multi-level stacking ensemble for detection of alzheimer's disease based on particle swarm optimization and the sub-scores of cognitive biomarkers," IEEE Access, 2023.
[31] T. Zhu, X. Liu, J. Wang, R. Kou, Y. Hu, M. Yuan, C. Yuan, L. Luo, and W. Zhang, "Explainable machine-learning algorithms to differentiate bipolar disorder from major depressive disorder using self-reported symptoms, vital signs, and blood-based markers," Computer Methods and Programs in Biomedicine, vol. 240, p. 107723, 2023.
[32] A. Gkantzios, C. Kokkotis, D. Tsiptsios, S. Moustakidis, E. Gkartzonika, T. Avramidis, N. Aggelousis, and K. Vadikolias, "Evaluation of blood biomarkers and parameters for the prediction of stroke survivors' functional outcome upon discharge utilizing explainable machine learning," Diagnostics, vol. 13, no. 3, p. 532, 2023.
[33] M. Liu, J. Zhou, Q. Xi, Y. Liang, H. Li, P. Liang, Y. Guo, M. Liu, T. Temuqile, L. Yang et al., "A computational framework of routine test data for the cost-effective chronic disease prediction," Briefings in Bioinformatics, vol. 24, no. 2, p. bbad054, 2023.
[34] N. Rahim, T. Abuhmed, S. Mirjalili, S. El-Sappagh, and K. Muhammad, "Time-series visual explainability for alzheimer's disease progression detection for smart healthcare," Alexandria Engineering Journal, vol. 82, pp. 484–502, 2023.
[35] F. Curia, "Cervical cancer risk prediction with robust ensemble and explainable black boxes method," Health and Technology, vol. 11, no. 4, pp. 875–885, 2021.
[36] N. Settouti and M. Saidi, "Preliminary analysis of explainable machine learning methods for multiple myeloma chemotherapy treatment recognition," Evolutionary Intelligence, pp. 1–21, 2023.
[37] S. Deshmukh, B. K. Behera, P. Mulay, E. A. Ahmed, S. Al-Kuwari, P. Tiwari, and A. Farouk, "Explainable quantum clustering method to model medical data," Knowledge-Based Systems, vol. 267, p. 110413, 2023.
[38] M. R. Hassan, M. F. Islam, M. Z. Uddin, G. Ghoshal, M. M. Hassan, S. Huda, and G. Fortino, "Prostate cancer classification from ultrasound and mri images using deep learning based explainable artificial intelligence," Future Generation Computer Systems, vol. 127, pp. 462–472, 2022.
[39] E. Pintelas, M. Liaskos, I. E. Livieris, S. Kotsiantis, and P. Pintelas, "Explainable machine learning framework for image classification problems: case study on glioma cancer prediction," Journal of imaging, vol. 6, no. 6, p. 37, 2020.
[40] C. Severn, K. Suresh, C. Görg, Y. S. Choi, R. Jain, and D. Ghosh, "A pipeline for the implementation and visualization of explainable machine learning for medical imaging using radiomics features," Sensors, vol. 22, no. 14, p. 5205, 2022.
[41] H. V. Nguyen and H. Byeon, "Prediction of ecog performance status of lung cancer patients using lime-based machine learning," Mathematics, vol. 11, no. 10, p. 2354, 2023.
[42] R. O. Alabi, M. Elmusrati, I. Leivo, A. Almangush, and A. A. Mäkitie, "Machine learning explainability in nasopharyngeal cancer survival using lime and shap," Scientific Reports, vol. 13, no. 1, p. 8984, 2023.
[43] S. Alkhalaf, F. Alturise, A. A. Bahaddad, B. M. E. Elnaim, S. Shabana, S. Abdel-Khalek, and R. F. Mansour, "Adaptive aquila optimizer with explainable artificial intelligence-enabled cancer diagnosis on medical imaging," Cancers, vol. 15, no. 5, p. 1492, 2023.
[44] C. Cordova, R. Muñoz, R. Olivares, J.-G. Minonzio, C. Lozano, P. Gonzalez, I. Marchant, W. González-Arriagada, and P. Olivero, "Her2 classification in breast cancer cells: A new explainable machine learning application for immunohistochemistry," Oncology Letters, vol. 25, no. 2, pp. 1–9, 2023.
[45] D. Song, J. Yao, Y. Jiang, S. Shi, C. Cui, L. Wang, L. Wang, H. Wu, H. Tian, X. Ye et al., "A new xai framework with feature explainability for tumors decision-making in ultrasound data: comparing with grad-cam," Computer Methods and Programs in Biomedicine, vol. 235, p. 107527, 2023.
[46] K. Wickstrøm, K. Ø. Mikalsen, M. Kampffmeyer, A. Revhaug, and R. Jenssen, "Uncertainty-aware deep ensembles for reliable and explainable predictions of clinical time series," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 7, pp. 2435–2444, 2020.
[47] M. Ganeshkumar, V. Ravi, V. Sowmya, E. Gopalakrishnan, and K. Soman, "Explainable deep learning-based approach for multilabel classification of electrocardiogram," IEEE Transactions on Engineering Management, 2021.
[48] K. Wang, J. Tian, C. Zheng, H. Yang, J. Ren, Y. Liu, Q. Han, and Y. Zhang, "Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and shap," Computers in Biology and Medicine, vol. 137, p. 104813, 2021.
[49] H. V. Nguyen and H. Byeon, "Prediction of out-of-hospital cardiac arrest survival outcomes using a hybrid agnostic explanation tabnet model," Mathematics, vol. 11, no. 9, p. 2030, 2023.
[50] G. V. Aiosa, M. Palesi, and F. Sapuppo, "Explainable ai for decision support to obesity comorbidities diagnosis," IEEE Access, 2023.
[51] L. P. Joseph, E. A. Joseph, and R. Prasad, "Explainable diabetes classification using hybrid bayesian-optimized tabnet architecture," Computers in Biology and Medicine, vol. 151, p. 106178, 2022.

[52] B. Lalithadevi and S. Krishnaveni, "Diabetic retinopathy detection and severity classification using optimized deep learning with explainable ai technique," Multimedia Tools and Applications, pp. 1–65, 2024.
[53] Y. Du, A. R. Rafferty, F. M. McAuliffe, L. Wei, and C. Mooney, "An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus," Scientific Reports, vol. 12, no. 1, p. 1170, 2022.
[54] Y. Du, A. R. Rafferty, F. M. McAuliffe, J. Mehegan, and C. Mooney, "Towards an explainable clinical decision support system for large-for-gestational-age births," Plos one, vol. 18, no. 2, p. e0281821, 2023.
[55] B. Lalithadevi, S. Krishnaveni, and J. S. C. Gnanadurai, "A feasibility study of diabetic retinopathy detection in type ii diabetic patients based on explainable artificial intelligence," Journal of Medical Systems, vol. 47, no. 1, p. 85, 2023.
[56] K. Debjit, M. S. Islam, M. A. Rahman, F. T. Pinki, R. D. Nath, S. Al-Ahmadi, M. S. Hossain, K. M. Mumenin, and M. A. Awal, "An improved machine-learning approach for covid-19 prediction using harris hawks optimization and feature analysis using shap," Diagnostics, vol. 12, no. 5, p. 1023, 2022.
[57] V. V. Khanna, K. Chadaga, N. Sampathila, S. Prabhu, and R. Chadaga, "A machine learning and explainable artificial intelligence triage-prediction system for covid-19," Decision Analytics Journal, p. 100246, 2023.
[58] L. Zou, H. L. Goh, C. J. Y. Liew, J. L. Quah, G. T. Gu, J. J. Chew, M. P. Kumar, C. G. L. 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, no. 2, pp. 242–254, 2022.
[59] M. Cavallaro, H. Moiz, M. J. Keeling, and N. D. McCarthy, "Contrasting factors associated with covid-19-related icu admission and death outcomes in hospitalised patients by means of shapley values," PLOS Computational Biology, vol. 17, no. 6, p. e1009121, 2021.
[60] K. Chadaga, S. Prabhu, V. Bhat, N. Sampathila, S. Umakanth, and R. Chadaga, "A decision support system for diagnosis of covid-19 from non-covid-19 influenza-like illness using explainable artificial intelligence," Bioengineering, vol. 10, no. 4, p. 439, 2023.
[61] M. Laatifi, S. Douzi, H. Ezzine, C. E. Asry, A. Naya, A. Bouklouze, Y. Zaid, and M. Naciri, "Explanatory predictive model for covid-19 severity risk employing machine learning, shapley addition, and lime," Scientific Reports, vol. 13, no. 1, p. 5481, 2023.
[62] F. Juraev, S. El-Sappagh, E. Abdukhamidov, F. Ali, and T. Abuhmed, "Multilayer dynamic ensemble model for intensive care unit mortality prediction of neonate patients," Journal of Biomedical Informatics, vol. 135, p. 104216, 2022.
[63] Z. Zeng, X. Tang, Y. Liu, Z. He, and X. Gong, "Interpretable recurrent neural network models for dynamic prediction of the extubation failure risk in patients with invasive mechanical ventilation in the intensive care unit," BioData Mining, vol. 15, no. 1, pp. 1–19, 2022.
[64] A. J. Barda, C. M. Horvat, and H. Hochheiser, "A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare," BMC medical informatics and decision making, vol. 20, no. 1, pp. 1–16, 2020.
[65] H.-C. Thorsen-Meyer, A. B. Nielsen, A. P. Nielsen, B. S. Kaas-Hansen, P. Toft, J. Schierbeck, T. Strøm, P. J. Chmura, M. Heimann, L. Dybdahl et al., "Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records," The Lancet Digital Health, vol. 2, no. 4, pp. e179–e191, 2020.
[66] S. Knapič, A. Malhi, R. Saluja, and K. Främling, "Explainable artificial intelligence for human decision support system in the medical domain," Machine Learning and Knowledge Extraction, vol. 3, no. 3, pp. 740–770, 2021.
[67] D. Mukhtorov, M. Rakhmonova, S. Muksimova, and Y.-I. Cho, "Endoscopic image classification based on explainable deep learning," Sensors, vol. 23, no. 6, p. 3176, 2023.
[68] D. Varam, R. Mitra, M. Mkadmi, R. Riyas, D. A. Abuhani, S. Dhou, and A. Alzaatreh, "Wireless capsule endoscopy image classification: An explainable ai approach," IEEE Access, 2023.
[69] N. Nigar, M. Umar, M. K. Shahzad, S. Islam, and D. Abalo, "A deep learning approach based on explainable artificial intelligence for skin lesion classification," IEEE Access, vol. 10, pp. 113 715–113 725, 2022.
[70] Z. J. Lo, M. H. W. Mak, S. Liang, Y. M. Chan, C. C. Goh, T. Lai, A. Tan, P. Thng, J. Rodriguez, T. Weyde et al., "Development of an explainable artificial intelligence model for asian vascular wound images," International Wound Journal, 2023.
[71] C. Metta, A. Beretta, R. Guidotti, Y. Yin, P. Gallinari, S. Rinzivillo, and F. Giannotti, "Improving trust and confidence in medical skin lesion diagnosis through explainable deep learning," International Journal of Data Science and Analytics, pp. 1–13, 2023.
[72] S. Martínez-Agüero, C. Soguero-Ruiz, J. M. Alonso-Moral, I. Mora-Jiménez, J. Álvarez-Rodríguez, and A. G. Marques, "Interpretable clinical time-series modeling with intelligent feature selection for early prediction of antimicrobial multidrug resistance," Future Generation Computer Systems, vol. 133, pp. 68–83, 2022.
[73] M. Cavallaro, E. Moran, B. Collyer, N. D. McCarthy, C. Green, and M. J. Keeling, "Informing antimicrobial stewardship with explainable ai," PLOS Digital Health, vol. 2, no. 1, p. e0000162, 2023.
[74] M. S. Islam, M. A. Awal, J. N. Laboni, F. T. Pinki, S. Karmokar, K. M. Mumenin, S. Al-Ahmadi, M. A. Rahman, M. S. Hossain, and S. Mirjalili, "Hgsorf: Henry gas solubility optimization-based random forest for c-section prediction and xai-based cause analysis," Computers in Biology and Medicine, vol. 147, p. 105671, 2022.
[75] I. K. Kokkinidis, E. Logaras, E. S. Rigas, I. Tsakiridis, T. Dagklis, A. Billis, and P. D. Bamidis, "Towards an explainable ai-based tool to predict preterm birth," CARING IS SHARING–EXPLOITING THE VALUE IN DATA FOR HEALTH AND INNOVATION, p. 571, 2023.
[76] K. H. Cho, E. S. Kim, J. W. Kim, C.-H. Yun, J.-W. Jang, P. H. Kasani, and H. S. Jo, "Comparative effectiveness of explainable machine learning approaches for extrauterine growth restriction classification in preterm infants using longitudinal data," Frontiers in Medicine, vol. 10, 2023.
[77] F. Curia, "Features and explainable methods for cytokines analysis of dry eye disease in hiv infected patients," Healthcare Analytics, vol. 1, p. 100001, 2021.
[78] R.-K. Sheu, M. S. Pardeshi, K.-C. Pai, L.-C. Chen, C.-L. Wu, and W.-C. Chen, "Interpretable classification of pneumonia infection using explainable ai (xai-icp)," IEEE Access, vol. 11, pp. 28 896–28 919, 2023.
[79] J. Peng, K. Zou, M. Zhou, Y. Teng, X. Zhu, F. Zhang, and J. Xu, "An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients," Journal of medical systems, vol. 45, pp. 1–9, 2021.
[80] H.-C. Chen, C. Damarjati, K. T. Putra, H.-M. Chen, C.-L. Hsieh, H.-J. Lin, M.-Y. Wu, and C.-S. Chen, "Pulse-line intersection method with unboxed artificial intelligence for hesitant pulse wave classification," Information Processing & Management, vol. 59, no. 2, p. 102855, 2022.
[81] S. M. Lauritsen, M. Kristensen, M. V. Olsen, M. S. Larsen, K. M. Lauritsen, M. J. Jørgensen, J. Lange, and B. Thiesson, "Explainable artificial intelligence model to predict acute critical illness from electronic health records," Nature communications, vol. 11, no. 1, p. 3852, 2020.
[82] K. Davagdorj, J.-W. Bae, V.-H. Pham, N. Theera-Umpon, and K. H. Ryu, "Explainable artificial intelligence based framework for non-communicable diseases prediction," IEEE Access, vol. 9, pp. 123 672–123 688, 2021.
[83] H. Zhang and K. Ogasawara, "Grad-cam-based explainable artificial intelligence related to medical text processing," Bioengineering, vol. 10, no. 9, p. 1070, 2023.
[84] K. Mridha, S. Ghimire, J. Shin, A. Aran, M. M. Uddin, and M. Mridha, "Automated stroke prediction using machine learning: An explainable and exploratory study with a web application for early intervention," IEEE Access, 2023.
[85] N. A. Wani, R. Kumar, and J. Bedi, "Deepxplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence," Computer Methods and Programs in Biomedicine, vol. 243, p. 107879, 2024.
[86] M. M. Hassan, S. A. AlQahtani, M. S. AlRakhami, and A. Z. Elhendi, "Transparent and accurate covid-19 diagnosis: Integrating explainable ai with advanced deep learning in ct imaging." CMES-Computer Modeling in Engineering & Sciences, vol. 139, no. 3, 2024.
[87] S. Liu, A. B. McCoy, J. F. Peterson, T. A. Lasko, D. F. Sittig, S. D. Nelson, J. Andrews, L. Patterson, C. M. Cobb, D. Mulherin et al., "Leveraging explainable artificial intelligence to optimize clinical decision support," Journal of the American Medical Informatics Association, vol. 31, no. 4, pp. 968–974, 2024.
[88] G. Schwalbe and B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts," Data Mining and Knowledge Discovery, pp. 1–59, 2023.
[89] M. A. Qureshi, A. Younus, and S. Caton, "Inclusive counterfactual generation: Leveraging llms in identifying online hate," in International Conference on Web Engineering. Springer, 2024, pp. 34–48.