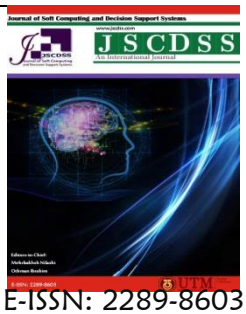
An International Journal
JSCDSS
http://www.jscdss.com
Vol.11 No.1 February 2024: 1-22
Article history:
Accepted 12 February 2024
Published online 12 February 2024

Journal of Soft Computing and Decision Support Systems



E-ISSN: 2289-8603

Does Explainability Enhance the Effectiveness of AI Models in Public Health? The COVID-19 Context

Neda Ahmadi [a], Mehrbakhsh Nilashi [b,c,*]

[a] Faculty of Engineering, Computing and the Environment, School of Computer Science and Mathematics, Department of Computer Science, Kingston University London, KT1 2EE London, United Kingdom

[b] UCSI Graduate Business School, UCSI University, No. 1 Jalan Menara Gading, UCSI Heights, 56000, Cheras, Kuala Lumpur, Malaysia

[c] Centre for Global Sustainability Studies (CGSS), Universiti Sains Malaysia, 11800 Penang, Malaysia

*** Corresponding author email address**: nilashidotnet@hotmail.com

Abstract

Generative AI models, such as ChatGPT, offer versatile applications in healthcare, particularly in the COVID-19 era. While these models show promise in medical decision support, the imperative of explainability cannot be overstated. Understanding how AI arrives at recommendations is crucial for transparency and trust, especially in critical areas like COVID-19 management. However, challenges persist in elucidating the decision-making processes of AI models, potentially hindering their acceptance in medical practice. This paper discusses the necessity of prioritizing explainability mechanisms tailored for AI-powered linguistic models, particularly in the context of COVID-19-related healthcare decisions. By shedding light on AI reasoning, explainability mechanisms not only enhance transparency and accountability but also foster trust among medical professionals, facilitating informed collaboration between human expertise and AI capabilities.

Keywords: Explainability, ChatGPT, COVID-19, XAI, Healthcare, Machine Learning

1. Introduction

In 2019, SARS-CoV-2, or COVID-19, first emerged in the city of Wuhan, China. Initially, it appeared as a localized occurrence of respiratory distress, attributed to a novel mutation of a previously unknown pathogen. COVID-19 has profoundly impacted people's lives worldwide (Abumalloh et al., 2021; Nilashi, Abumalloh, et al., 2021; Nilashi, Abumalloh, Alrizq, Alghamdi, et al., 2022; Nilashi, Abumalloh, Alrizq, Almulihi, et al., 2022; Nilashi, Abumalloh, Minaei-Bidgoli, Zogaan, et al., 2022; Nilashi, Abumalloh, Mohd, et al., 2023; Nilashi, Asadi, et al., 2021; Rupani et al., 2020; Taheri et al., 2021), causing widespread illness and mortality, and increased levels of stress and anxiety (Husky, Kovess-Masfety, & Swendsen, 2020). As of October 18, 2023, the World Health Organization (WHO) has recorded a total of 771,407,825 reported instances of COVID-19 worldwide, along with 6,972,152 reported fatalities (WHO, 2023). Billions of individuals have received COVID-19 vaccines globally, safeguarding them from the SARS-CoV-2 virus and preventing over 20 million fatalities (Callaway, 2023). However, certain viral variants have demonstrated the ability to partially elude the immunity conferred by the initial vaccines. Consequently, vaccine researchers worldwide are actively involved in the development of numerous 'next-generation' COVID-19 vaccines (Callaway, 2023). Like many other viruses, COVID-19 is susceptible to evolving over time, and this evolution can result from various factors, including environmental influences. While some of these changes are minor and do not substantially affect the virus's characteristics, others can have a significant impact, such as increasing its contagiousness. These alterations in the virus are referred to as mutations, and when one or more mutations occur, the virus is categorized as a variant. Whenever a new variant of the virus emerges, the WHO adopts a system of assigning it a letter from the Greek alphabet. In response to the emergence of new variants, the WHO and numerous

other healthcare organizations worldwide have implemented a classification system designed to categorize these variants based on their assessed risk to global public health.

Data science and artificial intelligence (AI) have been instrumental in combating COVID-19 from multiple angles. In public domain, Machine learning chatbots and data visualization tools have facilitated communication, education, and awareness. AI-driven contact tracing and outbreak detection have aided in early intervention. Employing machine learning (ML) algorithms enables the swift diagnosis and analysis of medical imaging, while natural language processing (Hall, Chang, & Jayne, 2022) helps process vast amounts of scientific literature for insights. The integration of data science and AI has yielded crucial resources for understanding, managing, and alleviating the effects of the epidemic at a worldwide level.

The recent expansion of ML systems can be attributed in part to advancements in hardware and other technologies. The broad adoption of sophisticated models, particularly deep neural networks (DNNs), is notable (Miikkulainen et al., 2024), and has played a significant role. However, the growing complexity of these models comes with a downside. Many of these systems function as black boxes , leaving users and those impacted with minimal comprehension of the underlying processes governing predictions (Hassija et al., 2023). This lack of transparency gives rise to various issues, including the potential for disastrous mistakes at the time of deploying inaccurate models or decision-making derived from them in practical, real-world situations (Gkontra, Quaglio, Garmendia, & Lekadir, 2023). Moreover, even when these systems prove successful, their inherent opacity can hinder acceptance within controlled sectors, legal frameworks, and the wider community as a whole (Vorm & Combs, 2022). Human nature tends to seek causation behind actions, possibly for valid reasons, leading to a reluctance to embrace techniques that lack direct interpretability, manageability, and reliability (Duque Anton, Schneider, & Schotten, 2022). This hesitation is especially pronounced given the escalating demand for ethical AI (Griffin, Green, & Welie, 2023). The field of medicine presents numerous formidable challenges that could be effectively tackled through the integration of AI. The critical nature of healthcare and the wealth of data from sources such as medical imaging, biosensors, molecular data, and electronic health records have propelled a surge in initiatives concentrating on the automation diagnostic process, prediction outcome, formulating drugs, and conducting analysis in recent developments (AlZubi, Alarifi, & Al-Maitah, 2020; Bashir et al., 2023; Javaid, Haleem, Singh, Suman, & Rab, 2022; Muniz et al., 2010; Pedrero-Sánchez, Belda-Lois, Serra-Ano, Ingles, & Lopez-Pascual, 2022; Qureshi et al., 2023; Sharma, Gulati, & Chopra, 2023; Tong et al., 2023; Y. Zhang, 2017). The overarching goals of AI in medicine encompass tailoring decisions in healthcare (Ahmadi, Gholamzadeh, Shahmoradi, Nilashi, & Rashvand, 2018; Ghane, Ang, Nilashi, & Sorooshian, 2022; Nilashi, Abumalloh, Minaei-Bidgoli, Samad, et al., 2022; Nilashi, Abumalloh, Yusuf, et al., 2023; Nilashi, Ahmadi, Shahmoradi, Ibrahim, & Akbari, 2019; Nilashi et al., 2020; Nilashi, Bin Ibrahim, Mardani, Ahani, & Jusoh, 2018; Nilashi, Ibrahim, Ahmadi, Shahmoradi, & Farahmand, 2018; Nilashi, Ibrahim, et al., 2019), practices for well-being, and personalized treatments for Individuals receiving healthcare (Ahmed, Mohamed, Zeeshan, & Dong, 2020; Amann et al., 2020). However, current status regarding machine intelligence in the medical domain is characterized as "high on promise and relatively low on data and proof" (Antoniadi et al., 2021; Topol, 2019). Several machine intelligence devices exhibit proficiency in practical situations, accomplishing tasks like diabetes-related retinal disease, wrist fracture identification, histopathological breast malignancy dissemination, and detection of congenital clouding of the eye lens. Nevertheless, despite demonstrating equivalence or superiority to experts in experimental settings, many of these systems exhibit excessive occurrence of incorrect positive identifications when deployed in actual clinical environments (Kelly, Karthikesalingam, Suleyman, Corrado, & King, 2019; Mehta, Liao, Jenkinson, Carneiro, & Verjans, 2022).

1.1 *Foundational Aspects of XAI*

In the examination of the comprehensiveness of AI models (Arrieta et al., 2020), three dimensions merit consideration:

- Explainability: This is an active characteristic of a learning model, allowing for a clear articulation of the processes it undergoes. The primary objective revolves around elucidating the internal workings concerning the learning framework. Notably, explainability is not pursued merely out of intellectual curiosity; its significance becomes paramount during the moments that When human lives are endangered. (Xie, Gao, & Chen, 2019).
- Interpretability: as opposed to explainability, is a passive attribute of a learning framework. It enables individuals to understand the model and extract meaning from that.
- Transparency: Transparency, intricately tied to comprehensibility, considers a learning

An International Journal JSCDSS E-ISSN: 2289-8603

framework transparent when the learning system demonstrates clarity autonomously, without depending on any external interface. Transparency is attained when a learning model is inherently comprehensible without the need for additional components. Notably, transparency encompasses both explainability and interpretability, emerging as a pivotal aspect in the holistic comprehension of a learning model. The scholarly and industrial discussions on comprehensibility within machine learning (ML) and deep learning (DL) frameworks have been fervent, with Tjoa et al. emphasizing its intrinsic advantages in improving the functioning of ML and DL frameworks. (Tjoa & Guan, 2020).

1.1.1 *Reasons of Demanding Explainable Artificial Intelligence*

Presently, we find ourselves surrounded by opaque machine intelligent frameworks that play a pivotal role in decision-making for various aspects of human life and society, such as self-driving cars, online communities, and healthcare infrastructure. Often, these decisions are made without a clear understanding of the justifications behind them.

As per (Adadi & Berrada, 2018) not most of the opaque machine intelligent frameworks require to elucidate the reason behind every choice as this may lead to numerous outcomes like diminishing performance effectiveness as well as enhancing production expenses. Broadly, clarity and comprehensibility are unnecessary in specific scenarios: firstly, Findings which are unsatisfactory are not along with serious outcomes; secondly, the issue has been thoroughly examined and rigorously validated in real-world situations, therefore, the determination taken via the opaque framework is reliable (Adadi & Berrada, 2018).

According to the acquired research in this study, the requirement for explainable machine intelligent can be explored through different viewpoints as follows:

Scientific viewpoint: In developing the opaque machine intelligent frameworks, our goal is to design a near equivalent to tackle the specific issue. Thus, following the development of the opaque machine intelligent framework, the developed framework signifies the core of understanding, instead of the information (Molnar, Casalicchio, & Bischl, 2020).

Industry viewpoint: Standards and the lack of individual confidence in opaque machine intelligence frameworks pose difficulties for businesses in utilizing sophisticated and precise systems. More transparent systems, which are easier to understand, are chosen in the business sector to comply with requirements. The significant benefit of explainable artificial intelligence is alleviating the typical compromise between model clarity and efficiency, thereby addressing these difficulties. Though, it may elevate expenses for the creation and implementation (Veiber, Allix, Arslan, Bissyandé, & Klein, 2020).

- From the evolutive viewpoint of models, the generation of improper outcomes in the opaque machine intelligent frameworks can stem from various factors, including restricted and prejudiced training data, the presence of anomalies, exposure to adversarial inputs, and issues related to overfit models. Understanding what these opaque machine intelligent frameworks are very important. Therefore, the focus is on leveraging these frameworks to comprehend, troubleshoot, and enhance the robustness of these frameworks (Doshi-Velez & Kim, 2017).

1.1.2 *Explainability and Interpretability*

It is noteworthy that terms explainability and interpretability may vary in definition across different literature sources (Adadi & Berrada, 2018; Arrieta et al., 2020; Das & Rad, 2020; Samek & Müller, 2019; Samih, Adadi, & Berrada, 2019). The crucial need for the transparency achieved through the implementation of explainability characteristics in machine intelligence frameworks and their decision-making processes is apparent. Concurrently, the increasing risks and hostile assaults on machine learning (ML) and deep learning (DL) algorithms emphasize the need for transparency in both computational and operational aspects of these systems (Akhtar & Mian, 2018; Ren, Zheng, Qin, & Liu, 2020). Clarity within AI models serves a diverse range of objectives, nurturing confidence in AI systems, assisting in the determination of whether to exclusively depend on AI or incorporate human elements in decision-making, and tackling security threats directed at AI-driven methodologies. A crucial aspect of XAI is responsibility, introducing a legal component into the framework. Upon enforcing the general data protection regulation (GDPR), the juridical dimension of responsibility gains greater prominence, particularly in nascent applications involving data of a privacy-aware nature, such as self-driving cars, the connectivity of devices, and immersive digital experiences. This highlights the augmented significance of legal scrutiny in the realm of evolving technologies. In compliance with GDPR, risk assessment becomes imperative to mitigate liability concerns. As an example, in healthcare IoT platforms collecting and aggregating information

An International Journal JSCDSS E-ISSN: 2289-8603

retrieved from body area network (BAN), processing information either on-site or in cloud systems. Disseminating outcomes to relevant parties requires careful consideration of potential accountability concerns in the event of unforeseen incidents. Considering these circumstances, the term explainable artificial intelligence (XAI) was coined, adding dimensions of clarity, openness, and responsibility to prevailing ML and DL (and consequently, machine intelligence) frameworks (Adadi & Berrada, 2018). XAI seeks to transform opaque machine intelligence frameworks into transparent methods, wherein the decision-making processes of algorithms and models are explainable and interpretable. Originating from a defense advanced research projects agency (DARPA) campaign, XAI bestows machine intelligence frameworks with traits akin to a transparent container. This facilitates intelligent devices in understanding their operational context and working collaboratively with humans in the decision-making process. Additionally, XAI aims to construct foundational frameworks to delineate practical occurrences, classified by DARPA as three groups: deep explanation, decipherable frameworks, and induction of models. In essence, two distinct systems are needed - one to interpret pre-existing intricate frameworks as well as another for elucidating recently crafted frameworks along with their decisions (Biran & Cotton, 2017).

1.2 *Explanatory Machine Intelligence Devices and Approaches*

Within the realm of previous works, diverse framworks emerge, aiming to elucidate existing ML models. Nevertheless, merely expounding on "how predictions are made" falls short of substantiating result validity. The limitations of the white-box model become apparent due to the intricate, quantitative, and non-intuitive nature of most models. Consequently, users lacking ML expertise find the model's functionalities inscrutable, regardless of its transparency. Thus, interpretable models are imperative for furnishing comprehensive insights into the choices and anticipations generated via the frameworks. Subsequently, we explore the two extensive categories of XAI systems.

1.2.1 Transparent Models

Termed interpretable models, these intentionally prioritize interpretability over black-box ML models. Explanation and understandability in these transparent models are approached through algorithmic transparency, followed by algorithmic decomposability and a model's simulation ability. These different types of explanations delineate the diverse levels of elucidation a specific model can offer. The initial level of transparency, algorithmic transparency, gauges a user's ability to understand the outcome derived from a specified collection of inputs. Precisely, a clear framework turns into entirely navigable through mathematical analyses and methods. Linear models lend themselves to easy interpretation, while nonlinear frameworks demand advanced techniques. The next tier regarding clarity revolves around a model's decomposability, gauging the sophistication of an elucidation for every component (input, assorted factors, as well as calculations) regarding a framework. This explanation facilitates a better grasp and articulation regarding how a framework behaves. Nevertheless, this necessitates each inputs to be easy understandable, a challenge for every model due to the potential lack of clarity in complex features and parameters, requiring additional tools for understanding. Moreover, complete transparency, or the third tier, is attained when a model is simulatable by humans. This level hinges on a model's ability to be simulated, with complexity taking center stage. To sum up, a framework that is easy to understand can be effortlessly clarified to individuals through text and visualizations. Additionally, a decomposable model, possessing simulation capability, implies a self-contained nature, allowing it to be comprehended, analyzed, and justified holistically by humans, without relying on supplementary tools.

1.2.2 *Anticipation Interpretation and Validation Frameworks*

Models designed as black boxes, lacking inherent interpretability, necessitate post-hoc explainability to augment their understanding through various explanation methods adopted by humans. These explicability methods encompass text-based explanations, visual representations, simplification-based explanations, illustrative examples, and feature relevance explanations. In a broader sense, the efficacy of post-hoc interpretability depends on the elucidation methods employed via interpreters, the categories regarding the information on which they are employed, as well as the specific methods in use. For example, interpreters may opt for explanation by simplification, coupled with conducting sensitivity analysis on images. Alternatively, interpreters might opt to provide explanations utilizing distinct representations or diverse illustrations. In (Doran, Schulz, & Besold, 2017), the authors delve into various types of black box models, offering a more extensive classification based on the transparency inferred by users. These systems

An International Journal JSCDSS E-ISSN: 2289-8603

fall into three categories: non-transparent, understandable, as well as transparent frameworks.
In non-transparent frameworks, the relationships between initial data and final results remain concealed from the user's view. In understandable frameworks, individuals not solely have knowledge of the input-to-output mappings but can also engage in mathematical analysis of these mappings. In transparent systems, individuals acquire understanding not just of the associations but also the particular principles dictating those associations. Essentially, transparent systems provide a comprehensive overview regarding associative representations (guidelines) alongside their corresponding outputs such a way that entirely understandable to individuals. Utilizing post-hoc interpretation methods allows for the derivation of an understandable framework from each opaque framework. The methods used to deduce an understandable framework from each opaque framework are additionally known as framework construction (Hagras, 2018). Nevertheless, interpreting the whole framework accompanied by a thorough elucidation remains challenging except there is a comprehensive portrayal of the framework. The consideration of local versus global context also proves to be a vital factor; explanations at the local level may not be entirely accurate in the global context (Ribeiro, Singh, & Guestrin, 2016).

1.3 *Parties Involved in XAI and Clarification Prerequisites*

As previously mentioned, placing complete trust in an intelligent system is contingent upon its ability to sufficiently justify the predictions it generates. Clients tend to have greater confidence in a system when it adheres to the caliber and explanation requirements set forth by the diverse stakeholders involved. Identifying these stakeholders stands out as a crucial initial step, followed by furnishing acceptable explanations and interpretations tailored to each recipient type, including engineers, designers, theorists, and others. Prior to examining the technical and arithmetic facets regarding XAI, the following conversation concisely investigates different participants and the necessary requirements crucial for crafting a coherent story. Once the numerical representation regarding the machine intelligence frameworks is accessible, assessing the functionality of the system becomes viable not just via numerical scrutiny, especially concerning particular characteristics, but also via appraising functional soundness. This involves feeding the framework with new information, acquiring the expected accurate result, thus confirming the precision of the framework in question. Although this engineering and mathematical viewpoint may not be essential to an ordinary customer, it holds paramount importance for system designers and developers. Their understanding of such a system is greatly enhanced when it is inherently transparent and interpretable by design, particularly during the developmental stages.

1.3.1 *Contributors to Explainable AI*

This section delves into the various contributors to explainable AI, elucidating their roles and expectations in the context of XAI.

i. End users and customers: Clarification holds vital significance for end users and customers to validate system operations and evaluate the accuracy of its results. Users demand reassurance via post-hoc elucidations, encompassing rules and ontology components, ensuring the framework operates accurately and delivers impartial outcomes. From the end user perspective, an interpretable AI model's clarification ought to consistently mirror individual psychological procedures as well as adhere to fundamental principles:
   - User comprehension of the theoretical framework as well as fundamental structure machines.
   - Visibility regarding the framework's complete operational efficiency to users, with ample explanations about its capabilities.
   - Intuitive mapping of visible system elements to respective functionalities.
   - Continuous user awareness of the present condition of the framework.

ii. Scholars and thinkers: Scholars focus on providing clear elucidations concerning openness as well as impartiality in machine intelligence frameworks, involving participants like computational experts, technologies, legal professionals, reporters, economists, and policymakers. Their objective is to move beyond technical details, ensuring fairness and unbiased behavior in AI systems to establish accountability. Ethicists demand assurance via post-hoc interpretations which the framework reaches unbiased ethical decision-making. On the contrary, theorists are primarily focused on grasping the theoretical foundations of machine intelligence and pushing its boundaries. These participants, including scholars from academic or industrial backgrounds, prioritize gaining insights into the clear correlation between inner model conditions and the resulting outcome.

An International Journal JSCDSS E-ISSN: 2289-8603

iii. Technologists and mathematical experts: Quantitative structures offer numerical depictions for both theoretical concepts and practical applications. Technologists utilize these structures to illustrate the functioning of recently developed or pre-existing frameworks. When demanding explanations for engineers and scientists, it goes beyond data description or high-level discussions about the system's decisions. They expect transparency-based connections to the model's internal states (traceability to any action/state), making these explanations more than just post-hoc rationalizations.

1.3.2 *Explanation Requirements Based on Roles*

Identifying the explainer's role is crucial - who are the explanations intended for? Advanced mathematicians, engineers, employees, or customers? A clear delineation for "Explainer" roles is imperative for enhancing the relevance and interpretability of machine intelligence structures to make them more pertinent and easily understandable. Additionally, a transparent understanding regarding "What" is needed to perform the "explanation" is vital. It is essential to discern whether the focus is on the framework or the underlying information. In the context of a system designed to recognize different cars in the highways, a data-centric interpretation might clarify that a particular image is identified as a car due to its resemblance to different image of a vehicle (utilizing KNN), or the final result is reached through evaluating multiple characteristics. In essence, the clarity and interpretability ought to be tailored to specific use-cases or stakeholders (Tomsett, Braines, Harborne, Preece, & Chakraborty, 2018).

1.3.3 *Understanding for Technologists and Technicians Regarding Explainability*

The arithmetical elucidation of any machine intelligence framework essentially involves discussing the consequent key aspects:

1) Goal: establishing the objectives and intent regarding the framework, ensuring a clear understanding regarding what machine intelligence framework's intended accomplishments and goals.
   • Extent: Articulating the boundaries and extent of the machine intelligence framework, addressing inquiries regarding its application constraints or intricacies.
2) Model framework: Furnishing an overview of the systemic framework and the comprehensive operational design. This encompasses detailing variables or goals and operational intricacies, including the governing principles, regulations, and foundational assumptions essential for seamless system monitoring.
3) Mathematical representation and modeling: Presenting clear and elucidative mathematical expressions that depict the model of the system. These expressions establish connections between parameters, fixed values, and limitations. The process of mathematical modeling commences with a comprehensive equation outlining the inputs and potential outcomes of the model. Subsequently, intricate sub-units and logical factors are detailed. These depictions encapsulate insights into intermediate outputs and combinations leading to the final model output. Algorithmic transparency is another important parameter, defined as the user's ability to comprehend the framework and ascertain how a particular outcome is derived for a provided input parameter. Furthermore, a comprehensive explanation regarding the training dataset is essential to enhance the transparency and interpretability of the framework's predictions. Language-based policies play a vital role in deciphering training the data by establishing conditions such as "if x1 is high, then y is high," enabling the model to be simulated and enhancing its explanability. The significance of features is the other critical factor influencing the ultimate outcome result as well as contributing to framework clarity. For instance, in a framework explanation where five characteristics (x1, x2, x3, x4, x5) are opted, the prominence of x1 surpasses that of the others. A minor variation in the value of x1 can have a substantial impact on the framework outcome.

1.3.4 *Explanation of Presently Utilized Machine Intelligence Frameworks*

In this context, we explore the interpretability of frequently employed machine learning algorithms such as decision tree (DT) (Nilashi, Abumalloh, Almulihi, et al., 2023; Nilashi, bin Ibrahim, Ahmadi, & Shahmoradi, 2017; Nilashi, Rupani, et al., 2019; Nilashi, Samad, et al., 2021), K-nearest neighbors (KNN), and Bayesian structures, all recognized for their transparency. Decision trees are hierarchical models utilized for making the decisions, are easily simulated and comprehended by users when small. However, increased size impedes full evaluation, making it a decomposable model. Bayesian models, probabilistic and represented as directed acyclic graphical models, provide clear relationships between features and the target. For KNNs, the model's

An International Journal JSCDSS E-ISSN: 2289-8603

complexity and detailed variable descriptions are intuitive for humans, making them easily replicable. These models are transparent and decomposable, with mathematical tools used for complex cases.

In conclusion, the emergence of SARS-CoV-2 as well as the subsequent global impact regarding Corona virus infection have prompted unprecedented efforts within the domain of computer science and artificial intelligence to address and mitigate the challenges posed by the pandemic. The staggering numbers reported through the World Health Organization underscore the urgency as well as magnitude of the situation, necessitating innovative approaches and technologies. As the world grapples with the evolving nature of the virus, the development and deployment of COVID-19 vaccines have been pivotal in preventing millions of fatalities. However, the adaptability of the virus, as demonstrated by emerging variants, has underscored the ongoing need for research and advancements in vaccine technology. The ever-evolving scenario has paved the way for the amalgamation of artificial intelligence and data science to combat the pandemic. The expansion of ML systems, particularly the adoption of complex models like DNNs, has marked a significant stride forward. Nevertheless, the inherent opacity of these advanced models raises critical concerns, ranging from potential errors in real-world scenarios to challenges in gaining acceptance from various stakeholders. The demand for ethical AI further intensifies the need for transparency, interpretability, and accountability in these systems. Within the healthcare sphere, the incorporation of artificial intelligence presents potential for automating medical assessments and determinations, including diagnostic processes. Yet, the present state of machine intelligence in medical field is recognized for being "abundant in promise but somewhat lacking in substantial data and conclusive evidence." Implementing artificial intelligence driven structures within clinical environments has revealed challenges such as high false positive rates, emphasizing the importance of refining these technologies for real-world applications. The exploration of XAI becomes paramount in addressing the limitations of black-box AI models. The dimensions of explainability, interpretability, and transparency form the foundation of XAI, providing a framework to make AI systems more understandable and accountable. The emergence of XAI systems and techniques, including transparent models and anticipation interpretation and validation frameworks, offers avenues to narrow the divide within intricate frameworks and individual understanding. Comprehending the responsibilities and anticipations of diverse participants in explainable Artificial intelligence, ranging from end-users and customers to ethicists as well as engineers, becomes crucial in shaping explanations that cater to diverse needs. The discussion of explanation requirements based on roles emphasizes the importance of tailoring explanations to specific audiences, aligning comprehensibility with use-cases and stakeholders. As the journey into XAI progresses, the challenges and opportunities laid out in this section underscore the multidimensional nature of the intersection between AI and healthcare. Finding equilibrium within technological progress and individual comprehension remains a central theme, underscoring the necessity in regard to continual collaboration, research, as well as moral deliberations in influencing the trajectory regarding AI in the realm of public health and beyond.

The rest of the article is structured as follows. Section 2 provides an in-depth examination regarding explainability in generative AI models unfolds, exploring the landscape of XAI and its distinguishing characteristics from traditional AI. This section also delves into various explanation types within generative AI. Section 3 conducts an in-depth examination of XAI in medicine, elucidating the transformative impact of generative models on medical disease evaluation, with a specific focus on COVID-19 applications. Section 4 engages in a discussion on the implications of explainability in healthcare, considering the benefits, drawbacks, and ethical considerations. The article concludes in Section 5, summarizing critical discoveries and highlighting the crucial role regarding explainable artificial intelligence in improving transparency as well as reliance, particularly within medical field applications. The conclusion contemplates the wider consequences in regard to future investigation and the conscientious implementation regarding generative AI models in healthcare.

2. Explainability in Generative Artificial Intelligence Frameworks

The frameworks based on generative intelligence machines are a class of AI models (Hacker, Engel, & Mauer, 2023) that capable of producing novel material in various fields. These models use various techniques for producing different materials that is often indistinguishable from human-created content. Generative AI models have been effective and useful in healthcare practices (Meskó & Topol, 2023). In the field of medicine, ChatGPT has demonstrated a passing score in the United States Medical Licensing Examination (Kung et al., 2023). These models are designed for improving patient communication, clinical documentation, medical research, and etc.

An International Journal JSCDSS E-ISSN: 2289-8603

Explainability has been an important issue in the ML context and AI models (Sun et al., 2022). It has been found that the lack of explainability in AI models significantly can hinder their adoption. This is mainly more critical in healthcare decision-making. Devoid of the capability to comprehend how machine intelligence achieves preventive measures, diagnosis, treatment plan, or prognosis, clinicians may hesitate to rely on these tools for critical decisions. Due to the often-opaque nature of machine intelligence frameworks, commonly referred to as "black boxes," comprehending the decision-making process leading to their conclusions can be challenging. When AI models cannot explain their reasoning, it becomes difficult to hold them accountable for errors or biases that may arise in their output. In addition, lack of transparency can raise ethical and legal concerns, particularly when AI is involved in critical healthcare decisions.

Numerous concerns envelop the integration of machine learning in healthcare section, encompassing challenges related to partiality, confidentiality, safeguarding, and openness, causality, transferability, informativeness, fairness, and confidence (Antoniadi et al., 2021). Given that decisions influenced by these systems impact human health, there exists an urgent requirement to comprehend the processes involved in decision-making (Araujo, Helberger, Kruikemeier, & De Vreese, 2020; Asan, Bayrak, & Choudhury, 2020), particularly in critical areas like life-altering outcomes resulting from disease diagnosis (Gaviglio et al., 2023) and in precision medicine, scientists necessitate extensive data acquired through the structure beyond a basic dichotomous forecast (Arora et al., 2019). To tackle such problems, explanations for outputs of AI models are deemed crucial. Therefore, explainability, along with related concepts like interpretability and transparency, has emerged as a central concern in the field of ML in medicine in recent years (Arrieta et al., 2020; Loh et al., 2022). Despite the proven utility regarding ML-based systems, there is a contention that widespread adoption in routine medical practice is unlikely without addressing these challenges, primarily through the provision of adequate justifications in regard to the decisions rendered via these machines (Antoniadi et al., 2021; Picardi, Hawkins, Paterson, & Habli, 2019; Rahmani et al., 2021). Unfortunately, the complexity of different applications introduces additional complications, as various situations typically demand distinct interpretability and explainability requirements, hindering the development of generalizable solutions (Ding, Abdel-Basset, Hawash, & Ali, 2022; Lisboa, Saralajew, Vellido, Fernández-Domenech, & Villmann, 2023).

2.1 *Explainable AI (XAI)*

The term XAI first appeared in (Lane, Core, Van Lent, Solomon, & Gomboc, 2005). XAI is essentially about enhancing the comprehensibility of AI systems for individuals. Nevertheless, it is a lack of universally established methodological characterization for explainable machine intelligence, highlighting the need for increased clarity and consistency in the terminology used (Ali et al., 2023; Arrieta et al., 2020; Bellucci, Delestre, Malandain, & Zanni-Merk, 2021; Lopes, Silva, Braga, Oliveira, & Rosado, 2022). One challenge lies in the interchangeable deployment regarding the terms transparency, clarity, and explicability, despite their distinct meanings (Arrieta et al., 2020; Graziani et al., 2023; Saeed & Omlin, 2023). Interpretability pertains to the extent to which a model can be understood (Chakraborty et al., 2017), though it is sometimes used interchangeably with the term "explainability" (Erasmus, Brunet, & Fisher, 2021). Transparency can either denote a comprehensive attribute of "providing stakeholders with relevant information about how the model works," encompassing documentation regarding the training procedure, examination the arrangement of training data, software deployments, and explanations at the feature granularity (Jiang, Kahai, & Yang, 2022), or a precise algorithmic elucidation of the framework's functioning, as opposed to lack of transparency (Kemper & Kolkman, 2019; Langer & Landers, 2021). Explainability provides insights into the rationale behind the system's decision-making but is occasionally associated with understandability, which, by consensus, is loosely defined as "tools that empower a stakeholder to understand and, when necessary, contest the reasoning of model outcomes" (Antoniadi et al., 2021).

AI systems labeled as "black boxes," which provide predictions without offering any explanation, pose numerous issues. Beyond their lack of transparency, these systems conceal potential biases within their operations (Carabantes, 2020). Instances of balance in machine intelligence-driven forecasting platforms have been identified, contributing to the reinforcement of societal and classical biases. People who are historically excluded or marginalized in public bear disproportionate negative impacts (Packin & Lev Aretz, 2018). Predictive software utilized in legal settings to assess recidivism likelihood has demonstrated extreme unreliability, displaying biases based on race and assigning higher scores to Black individuals (Dressel & Farid, 2018). The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) AI system, specially, has faced widespread criticism for its unreliability and racial bias

An International Journal JSCDSS E-ISSN: 2289-8603

(Bagaric, Svilar, Bull, Hunter, & Stobbs, 2021). Numerous case studies reveal that the utilization of "dirty data" through platforms for policy-making leads to biased projections (Antoniadi et al., 2021). Search engines exhibit biases, favoring specific sites and revealing political biases (Bozdag, 2013). Algorithms used in employment, influenced due to "societal noise," sustain biased conduct that affect specific human as well as other associations (Ajunwa, 2020). This discriminatory pattern extends to focused promotions, with observed gender bias in the presentation of STEM job advertisements (Dalenberg, 2018). Additionally, vision detection systems demonstrate biases based on subject skin tone, exhibiting enhanced effectiveness for lighter Fitzpatrick complexions (Howard, Sirotin, Tipton, & Vemury, 2021). These examples highlight how these platforms, when implemented in practical scenarios, can transform into "Weapons of Math Destruction," perpetuating and exacerbating disparities (Carter, 2018).

Comparable challenges emerge as applications constructed on prejudiced information are employed in accuracy medicine (Habuza et al., 2021). Many medical datasets lack diversity, resulting in the creation of prejudiced frameworks. For example, populations in Europe are over-represented in genomic studies in the US, while other races are underrepresented (Landry, Ali, Williams, Rehm, & Bonham, 2018). Biases are evident in tools like the Framingham Heart Risk operations, which overestimate the likelihood of coronary heart sicknesses for certain populations due to the initial study's reliance on a predominantly Caucasian sample (D'Agostino Sr, Pencina, Massaro, & Coady, 2013). Implementing machine intelligence-driven framwrorks constructed using data in medicine field requires particular attention (Balagurunathan, Mitchell, & El Naqa, 2021). Despite the potential benefits of explanations in increasing understanding and trust in a system (Naiseh, Al-Thani, Jiang, & Ali, 2023), existing XAI techniques may not effectively identify discriminatory behavior in sensitive applications (Vale, El-Sharif, & Ali, 2022). While explanations can enhance understanding, simple explanations might conceal undesirable attributes and misguide individuals into drawing perilous or baseless finding, potentially leading to unethical outcomes (Gilpin et al., 2018). Awareness of the risks associated with blindly accepting justifications that could conceal ethnic or gender bias or promote a false perception of ethical values (fair-washing) (Alikhademi, Richardson, Drobina, & Gilbert, 2021) is crucial in all medical areas where such systems are employed.

2.2 *Explanation Types*

Explanatory methods could be categorized according to their extent, differentiating between methods furnishing system-wide clarification in regard to the whole platform and methods offering detailed insights for individual forecasts. Global explanations aid in understanding the overall reaction and rationale resulting in anticipated results. Conversely, individualized clarifications focus on justifying the specific decision for a particular instance (Kök, Okay, Muyanlı, & Özdemir, 2023). Furthermore, methods could be categorized into pre-modeling and post-modeling interpretability approaches. pre-modeling approaches, also known as clear approaches, are inherently designed to be explainable. These model-specific methods encompass some of the ML algorithms. Despite the variety of approaches, the aforementioned approaches have limits to their interpretability, especially in situations involving high dimensions and intricate complexities (Antoniadi et al., 2021). Some of the ML algorithms are inherently non-explainable within realistic boundaries and are often termed "black-box" models (Emmert-Streib, Yli-Harja, & Dehmer, 2020). post-modeling approaches, commonly independent of the specific model, might not elucidate the inner workings of opaque frameworks; however, they can provide localized interpretations in regard to the particular decisions (Vilone & Longo, 2021). One approach involves constructing clear and straightforward frameworks which offer understandable representations of the opaque frameworks (Guidotti et al., 2018).

Post-modeling approaches classify as overarching interpretations, such as the model-independent approach BETA (Mahya & Fürnkranz, 2023), neural network approach like GAM (Park, 2022), individualized clarifications, such as LIME (Natesan Ramamurthy, Vinzamuri, Zhang, & Dhurandhar, 2020), SHAP (Guillemé, Masson, Rozé, & Termier, 2019), as well as Anchors (Ribeiro, Singh, & Guestrin, 2018). These approaches furnish explanations at the feature level via constructing explainable models which mimic the action regarding the initial framework. System-wide interpretations could similarly be derived from these approaches through synthesizing individualized interpretations, such as SHAP overview visualizations (Melis, Kaur, Daumé III, Wallach, & Vaughan, 2021) or SP-LIME (Das & Rad, 2020). Model-agnostic methods like CLEAR (Ribeiro et al., 2018) and CERTIFAI (Udeshi et al., 2022) generate individualized explanations supported by counterfactual explanations, illustrating instances where the model yields a different outcome for inputs close to the original input. Various methods, like

An International Journal JSCDSS E-ISSN: 2289-8603

gradient attribution approaches (Rao, Böhle, & Schiele, 2022), deconvolution (McNally, Karpova, Cooper, & Conchello, 1999), and visual explanation techniques, contribute to the interpretability of different ML models. Evaluating post-hoc methods involves two main approaches: mathematically quantifiable metrics and human-centered evaluations (Lopes et al., 2022). However, consensus is lacking on how to assess interpretability, correctness of explanations, or benchmarking methods against each other (Bodria et al., 2023; X. Li et al., 2022). Concerns exist regarding the reliability for post-hoc explanations (Kenny, Ford, Quinn, & Keane, 2021) and the potential exposure of original models to adversarial attacks (Akhtar, Mian, Kardan, & Shah, 2021). Adversarial attacks can manipulate ML algorithms, influencing output with minimal changes to input data. Proposed solutions include applying explainable techniques like SHAP (Bhattacharya, 2022). Evaluation methods like "goodness checklist," scales for evaluating interpretation contentment, techniques for eliciting cognitive frameworks, and algorithmic metrics for assessor faithfulness, credibility of the interpretation, and the dependability of the framework have been proposed (Bhattacharya, 2022).

3. Explainable Artificial Intelligence in Healthcare

The integration regarding explainable machine intelligence in the medical field stems from the increasing need for healthcare scientists to grasp the reasoning beyond decisions made by machines. The demand is for AI approaches that not only exhibit high performance but also possess traits such as trustworthiness, transparency, interpretability, and explainability for human experts. This is not only crucial for medical professionals but also holds meaningful consequences for the society, regulations, and administration, because the clarity of AI equipment boosts the confidence of healthcare experts (Sheu & Pardeshi, 2022; Tjoa & Guan, 2020). Medical professionals often serve as intermediaries for individuals, conveying intricate information which patients may find challenging to understand and act upon. Clinical Decision Support Systems (CDSS) (Canny et al., 2023) that assist medical professionals in this translation process are valuable, provided they facilitate rather than impede this communication. Effectively interpreting the machine learning frameworks into healthcare demands building confidence among clinicians. However, instances like the erroneous recommendations made by "Watson for Oncology" (IBM) and misjudgments by a pneumonia prediction ML-based system have eroded confidence in the utilization regarding machine intelligence in healthcare (Tucci, Saary, & Doyle, 2022). The lack of understanding behind AI predictions in real-world clinical environments poses risks, emphasizing the need for explainable models to prevent potential mistakes (Yang, 2022).

Despite the growing recognition of the importance of explanations, there is a dearth of agreement on practical clarifications in varied environments (Antoniadi et al., 2021). Some innovative XAI approaches, such (Arrieta et al., 2020), encompass gradual result assistance in regard to healthcare diagnosis through the utilization of decision trees. This strategy enables machine intelligence framework to cooperate with individual professionals, facilitating joint decision-making. Wu et al. suggest an approach where experts are involved in the explanation process for convolutional neural networks (CNNs), demonstrating the potential for AI systems to produce explanatory descriptions supporting final classification decisions, particularly in the classification of diseased tissue (Wu et al., 2018). Creators of AI-driven medical systems are giving growing importance to interpretability. Zheng et al. (Zheng, Delingette, & Ayache, 2019) present a new and interpretable approach for classifying heart-related abnormalities, achieving 95% classification accuracy with transparency and trustworthiness. Tosun et al. (Tosun et al., 2020) explain HistoMapr-Breast, the ML-based tool utilizing in regard to biopsies from breast core, emphasizing the importance of integrating XAI systems into pathology workflows to tackle concerns related to prejudice, openness, security, and causation. They emphasize which the explainable machine intelligence framework should augment histopathologists, not replace them. Hicks et al. (Hicks et al., 2018) introduce Mimir, the automatic mixed-media documentation tool that enhances an explainability of deep neural network structures in healthcare issues. Mimir produces organized and meaningfully accurate documentations, enabling investigation as well as understanding of the DL algorithms' decision processes. Improved explanations lead to patient understanding and confidence in the logical progression, ultimately enhancing trust and enabling doctors to provide better diagnoses (Bussone, Stumpf, & O'Sullivan, 2015).

3.1 *AI Advancements in Medical Disease Evaluation*

In this part, we furnish a synopsis regarding the medical domain, focusing on recent developments in employing machine intelligence or XAI for addressing various infections and sicknesses. The research in the medical field predominantly relies on mathematical statistics and ML algorithms. This creates numerous opportunities for the implementation of XAI-based

An International Journal JSCDSS E-ISSN: 2289-8603

approaches to enhance current understanding through improved classification evaluation. One significant area of concern is severe respiratory distress disorder, where assessment is supported by different characteristics like critical indicators and thoracic radiographs (Selvaraju et al., 2017). The classification within this context can be efficiently conducted by combining or independently utilizing data from vital signs. Typically, a sick person exhibiting specific symptoms like cough and fever need to be initially identified, after which medical examiners employ vital signs and thoracic radiographs to diagnose and monitor the progress of pneumonia. Furthermore, this work facilitates predicting discharge, and a more detailed configuration can provide insights into algorithm behavior.

In terms of the classification mechanism, local decisions are determined by a single decision system, while global decisions involve multiple determinations. This nuanced approach helps in understanding and interpreting the intricate dynamics of the algorithm. The "ante-hoc" explanation type applies to models that are easily understood by humans, whereas "post-hoc" is relevant for black boxes and DNN. Among patients, bloodstream infection (BSI) is a frequently encountered condition, identifiable by the existence of bacteriological or mycotic microorganisms in blood specimens (Burnham, Rojek, & Kollef, 2018). Commonly referred to as bloodstream infection, BSI manifests intense indicators, including fever, elevated heart rate, hypertension, shivering, and digestive problems. Prior investigations explored bloodstream infection extensively, utilizing vital signs, laboratory variables, and intensive care units (ICU) admission data. Preprocessing primarily focuses on recovering unavailable information in both bloodstream infection and non-BSI instances, subsequently evaluated utilizing ML models. Detection of BSI allows for subsequent treatment with medication to facilitate recovery. Acute kidney injury (Lee et al., 2018) refers to the impairment of renal blood circulation or unfavorable repercussions of pharmaceuticals. Symptoms are typically evident in lab tests, including urinary production and blood creatinine concentrations. In cases involving breathing assistance, supplementary variables are regarded as parameters. Pre-processing is very important to augment data precision as well as yielding promising results. ML has proven effective in identifying the stage and severity of acute kidney injury (AKI), facilitating appropriate medication and treatment for mild cases while managing severe conditions. Predicting hospital mortality becomes particularly significant in comorbid or critical cases, involving distinct intensive care unit (ICU) parameters and medication-related effects. The initial filter for preprocessing focuses on critical cases, and data imputation is added as necessary. Previous works have made forecasts using temporal intervals, like preceding 48 or 72 hours.

The following literature encapsulates comprehensive examination for interpretation-focused pre-processing in medical AI. Various references, mechanisms, data preprocessing, evaluation methods/algorithms, and outcome/explanation types are presented. For instance, Selvaraju et al. (2017) utilized Grad-CAM for global explanations, employing visual geometry group, structured convolutional neural network, and reinforcement learning for correlations, yielding written clarifications (Selvaraju et al., 2017). In the context of researching treatments for patients affected by bloodstream infections (BSI), several studies are presented, each offering distinct insights. Burnham et al. (2018) explored a dataset of 430 individuals, employing some examinations and evaluation approaches. Multivariate Cox frameworks were utilized, and outcomes were measured through the ante hoc explanation approach (Burnham et al., 2018). Fabre et al. (2019) investigated 249 individuals, utilizing other tests and regression modeling with localized focus and multiple variables to estimate inclination metrics was implemented, assigned weights based on the reciprocal of the inclination metrics, adopting an ante-hoc explanation approach (Fabre, Amoah, Cosgrove, & Tamma, 2019). Harris et al. (2018) analyzed 391 individuals, utilizing Charlson comorbidity index (CCI) scores and multiple imputation. The binary modeling was employed, adopting a global explanation approach (Harris et al., 2018). Delahanty et al. (2018) worked with a substantial dataset over two million individuals, utilizing fivefold cross validation. The outcomes were assessed using some metrics such as sensitivity, specificity, and area under ROC and so on, adopting a post-processing explanation approach (Delahanty, Alvarez, Flynn, Sherwin, & Jones, 2019). Kam et al. (2017) focused on 5789 individuals, handling missing data and segmentation. Multi-layer perceptron, recurrent neural network, as well as long short-term memory structures were utilized; then, accuracy and area under ROC were assessed in a post-processing manner (Kam & Kim, 2017).

3.2 *XAI in COVID*-19

Since its emergence in December 2019, COVID-19, also recognized as the Coronavirus-2019, experience a global epidemic characterized by rapid development

An International Journal JSCDSS E-ISSN: 2289-8603

and an unprecedented rate of spread. The virus exhibits a high level of contagiousness and can manifest as asymptomatic, remaining mostly latent. However, it has the potential to progress swiftly, leading to often resulting in lethal respiratory illness in two to eight percent regarding those impacted individuals (Ye, Xia, & Yang, 2021). Defining the overall death rate, occurrence, and intensity of the infectious viral condition has proven challenging, in part due to the complexities linked to SARS-CoV-2 disease acquisition. Factors contributing to this complexity include the escalation in the quantity of viral particles around the onset of symptoms or just before, as well as a misunderstanding of the abnormal physiological processes involving multiple organs, with prominent characteristics and a tendency for lung-related fatalities may be misconstrued. The proliferation of cases has placed immense strain on medical infrastructure globally, exacerbated by lack of essential safeguarding measures and specialized healthcare providers, further compounded by the absence of economical diagnostic procedures. Even with an advent of the rapid reverse transcription polymerase chain reaction (RT-PCR) test, difficulties persist, including elevated rate of false-negative results, delayed processing durations, variability in testing procedures, reported identification accuracy as limited as 60 to70 percent (Hu et al., 2020). As of May 17, 2021, there have been over 160 million reported cases of COVID-19 infection and over three million fatalities globally. In situations of scarcity, ethical considerations dictate which medical facility assets need to be allocated with priority regarding the individuals experiencing the most severe impacts of the condition.

Multiple recent studies in the domain of medical image analysis indicate that ML models significantly enhance computer-aided medical diagnosis for images obtained through thoracic radiographs as well as computed tomography (CT). The aforementioned representations prove beneficial for disease detection, and encompassing conditions. Moreover, various international studies, outlined in the literature, underscore the significance of employing DL algorithms for swift COVID-19 diagnosis through medical image datasets (Ozyurt, Tuncer, & Subasi, 2021; Thakur & Kumar, 2021; Tuncer, Ozyurt, Dogan, & Subasi, 2021; J. Zhang et al., 2021). Much of the research has demonstrated robust categorization efficacy utilizing DL methodologies. As an example, Ozyurt et al. suggested the combination of attribute generator and a repetitive combined attribute extractor employing a four-step image pre-processing method extracting manually designed attributes from CT images. An ANNs and DNN structures utilized such attributes for categorizing both in good health and non-healthy images, achieving accuracy in categorization about 94.10% percent as well as 95.84 percent (Ozyurt et al., 2021). In another study (Taresh, Zhu, Ali, Hameed, & Mutar, 2021), Zhu et al. emphasized pre-existing trained methods designed for the categorization of thoracic radiograph images. Nonetheless, it's noteworthy that not any of the scholarly articles placed emphasis on interpretability of the framework.

To be trustworthy, generative AI models must be understandable. They must be transparent and interpretable in order to be understandable, and this is an operationalized approximation for explainability. Despite generative AI models' great potential in the COVID-19 context, its drawbacks and limitations should not be neglected. Explainability is yet an obstacle for current generative AI models applications in medical decisions for COVID-19 management in different aspects such as public health education, symptom assessment, vaccination information research and insights and so on. Still, the recommendations of experts in medical science are inevitable for final decisions. Thus, generative AI models' lack of XAI can be a critical limitation in its wide adoption in the COVID-19 context, especially in sensitive aspects like diagnosis and treatment decisions. Without XAI, it can be challenging to fully comprehend and justify the responses of the model, which is a fundamental requirement in the process of making medical decisions.

In conclusion, the rapid global spread of Coronavirus has strained medical infrastructures, prompting the enhancement regarding AI well-being informatics resolutions. Notably, ML models, particularly DL structures, have demonstrated significant potential within diagnosing Coronavirus through diagnostic visualization. However, the lack of emphasis on model explainability in existing research poses challenges to their widespread clinical application. Addressing this limitation is crucial, as the clarity and comprehensibility of machine intelligent frameworks, especially within the generative AI deployments, are vital concerning reliable and well-informed decision processes in medical contexts, like diagnosis and treatment decisions related to COVID-19. Incorporating XAI techniques is pivotal to ensure the dependability as well as acceptance regarding machine intelligent frameworks in critical healthcare scenarios, ultimately enhancing their utility in pandemic management.

An International Journal
JSCDSS
E-ISSN: 2289-8603

4. Discussion

Explainability in generative machine intelligent frameworks within the domain of Coronavirus holds immense importance from various perspectives, including patients, healthcare managers, physicians, and researchers. For COVID-19 patients, the capability to understand the rationale underlying AI COVID-19 recommendations or diagnoses is essential for informed decision-making and trust-building. When patients understand why a particular COVID-19 treatment plan or preventive measure is suggested by an AI model, they are more likely to adhere to it. This, in turn, contributes to better COVID-19 health outcomes and patient satisfaction. Healthcare managers and administrators dealing in the challenges regarding the Coronavirus benefit from machine intelligent model explainability as it aids in resource allocation, policy development, and system optimization. Understanding why AI recommends certain actions during the COVID-19 crisis enables healthcare leaders to make data-driven decisions about COVID-19 staffing, equipment procurement, and facility management. Explainable AI also helps in evaluating the effectiveness of COVID-19 healthcare policies and interventions, enabling adjustments as needed for more efficient resource utilization and better overall COVID-19 pandemic management.

Poor generalization and the predominant reliance on shortcut learning characterize ML models constructed and educated in the manner of latest research. These models, constructed akin to recent studies, exhibit poor generalization, primarily relying on the acquisition regarding the abbreviations. The aforementioned undesirable action is, to some extent, connected to the combination of data used for training purposes derived from separate data sets containing images from both Coronavirus negative and positive cases, introducing significant confounding and creating ample opportunities for shortcut learning. Notably, the study warns that relying solely on validation using an independent evaluation set might not be adequate for identifying inadequately performing frameworks, as shortcuts may persist across internal and external domains. Earlier investigations into machine intelligent platforms regarding Coronavirus identification in X-ray images have shown mixed success in recognizing shortcuts. Noteworthy is a study demonstrating that models maintain high performance by focusing solely on the edges or boundaries of X-ray images, effectively removing authentic Coronavirus abnormality (Maguolo & Nanni, 2021). While this study aligns with the current findings, it primarily discusses the potential issue rather than its real-world occurrence. Another study using the GSInquire approach (Wang, Lin, & Wong, 2020) For saliency map analysis, there are no indications of a simplified learning process in the trio of released visuals. However, considering the likeness between its training dataset and the present dataset, as well as the observed deviation in effectiveness, suspicions arise regarding the possibility of shortcut learning. The discussion recommends that researchers employ interpretable machine intelligent or the methods based on prominence diagram at the level of the entire populace, emphasizing the need for caution in interpreting high performances without external validation (Ghoshal & Tucker, 2020; Karim et al., 2020). The study underscores the labor-intensive nature of valuations at the scale of the entire populace utilizing prominence diagrams and suggests the necessity for upcoming strategies in interpretable machine intelligence in the realm of diagnostic scans whereby streamline populace-wide assessment. Practical measures to alleviate simplified learning process are proposed, including improved collection of training data to align with the target population and the incorporation of clinically relevant labels. Despite these measures, the study highlights that simplified learning process might persist even under optimal data acquisition scenarios, underscoring an ongoing need for caution and principled external validation in the enhancement as well as acceptance regarding machine intelligent approaches for radiographic COVID-19 detection (M. D. Li et al., 2020).

Physicians, who play a central role in COVID-19 patient care, rely on AI models to aid in COVID-19 diagnosis, treatment planning, and patient monitoring. Explainability is important in regard to establishing the trust for medical practitioners in COVID-19 AI recommendations (Fuhrman et al., 2022). When doctors understand the rationale behind AI-generated COVID-19 suggestions, they can make more confident and informed clinical decisions. This collaborative approach between physicians and AI technologies enhances the quality of COVID-19 care, reduces diagnostic errors, and optimizes COVID-19 treatment plans, ultimately benefiting COVID-19 patient health. Generative AI models are widely used in education (Peres, Schreier, Schweidel, & Sorescu, 2023) and researchers working on COVID-19-related studies and clinical trials also benefit from AI model explainability. Transparent AI models provide insights into the COVID-19 features and data patterns that contribute to certain outcomes, aiding researchers in uncovering potential new avenues of COVID-19 investigation. Additionally, explainability can help researchers identify data quality issues, biases, or anomalies in their COVID-19 datasets, ensuring that

An International Journal JSCDSS E-ISSN: 2289-8603

studies are based on reliable and unbiased COVID-19 information. Therefore, incorporating XAI into generative AI models is highly beneficial as it influences the process of resolution regarding machine intelligent frameworks that are clearer and comprehensible to humans. In the context of COVID-19, XAI techniques provide clinicians and individuals with clear insights into how the model arrives at medical recommendations, ensure transparency and trust. In addition, integrating visualization technologies for explanations in generative AI models by combining textual outputs with visual outputs will aid to enhance the clarity and openness regarding the machine intelligent-generated responses. Within the context of text inputs, the generation of attention maps can prove to be beneficial as they serve the purpose of highlighting the specific words or phrases within the input text that have garnered the model's attention during the process of generating the response. The inclusion of these maps in conjunction with the generated response serves to provide a visual representation of the cognitive processes employed by the model. When providing a diagnosis based on COVID-19 patient symptoms, highlighting the symptoms in the input text by attention map that most influenced the diagnosis can be significant. This map can be displayed alongside the diagnosis, visually indicating the reasoning behind the decision. For dialogue-based interactions, it is important to display a timeline of the conversation along with visual markers indicating which parts of the conversation influenced each response.

XAI has become a potent instrument in combating the Coronavirus pandemic, especially through the utilization regarding generative models. These models excel in deciphering intricate patterns within medical data, facilitating the swift and accurate identification of COVID-19. In diagnostic applications, XAI offers medical practitioners providing transparent understanding regarding machine intelligence frameworks' decision procedure, enhancing their trust as well as confidence in the technology. The interpretability of generative models allows clinicians to understand the features contributing to a particular prediction, aiding in the validation of results and the optimization of treatment plans. Furthermore, the ability of XAI, especially generative models, to generate synthetic data has proven invaluable in situations where access to large, labeled datasets is limited. This adaptability is particularly advantageous in the dynamic landscape of COVID-19, where the virus's evolution necessitates rapid model adjustments and enhancements. An example of the benefits of XAI, specifically generative models, in COVID-19 applications is the synthesis of medical images. Generative adversarial networks (GANs) can be utilized to create authentic radiographs of the chest or CT scans depicting COVID-19 patients. Synthetic data augmentation helps overcome limitations in real-world datasets, especially in scenarios where having a substantial quantity of tagged COVID-19 images to use in training is challenging. By creating additional diverse examples, generative models enhance the robustness of AI algorithms, improving diagnostic accuracy and contributing to the development of more effective COVID-19 detection systems.

Despite the promising benefits, the integration of XAI in COVID-19 applications raises critical challenges and ethical considerations. One notable drawback is the potential for biased predictions, as machine intelligence might unwittingly sustain existing healthcare disparities available within the training set. Ensuring fairness in AI models demands concerted efforts to identify and rectify bias, emphasizing the need for diverse and representative datasets. Ethical concerns extend to patient privacy and data security, with the use of sensitive medical information to train generative models prompting questions about consent, confidentiality, and data ownership. Achieving a delicate equilibrium in the clarity regarding XAI and protecting individual confidentiality is paramount, necessitating the establishment involving strong ethical principles and guidelines. This responsible implementing regarding explainable artificial intelligence in medical fields needs continuous cooperation among technologists, medical practitioners, as well as policy makers for addressing these challenges and ensure the ethical use regarding machine intelligent tools employed in combating the Coronavirus pandemic. Consider a situation where an XAI model is utilized in regard to Coronavirus forecasting the likelihood of adverse events according to the patient health records. If the training data predominantly includes information from a specific demographic group, such as a particular ethnicity or socioeconomic status, the model may inadvertently introduce bias. For instance, if the framework underwent training using data predominantly derived from city dwellers; this might not generalize well to rural areas, potentially exacerbating healthcare disparities. Ethical concerns come to the forefront in scenarios where AI is applied to contact tracing. If the system collects extensive personal information, such as location data, without clear consent or robust anonymization measures, it could violate individual privacy rights and lead to unintended consequences, emphasizing the need for stringent ethical guidelines and safeguards.

An International Journal JSCDSS E-ISSN: 2289-8603

5. Conclusion

In conclusion, this extensive exploration into the role of explainability in generative AI models within the challenging landscape of healthcare underscores the critical importance of transparency, accountability, and ethical considerations. The multifaceted nature of this discussion has illuminated how explainability profoundly impacts various stakeholders, including patients, healthcare administrators, physicians, and researchers, especially within the framework regarding the ongoing battle against the global Coronavirus. From a patient-centric perspective, the comprehension of AI-generated recommendations not only fosters trust but also empowers individuals to actively participate in their healthcare decisions. This heightened level of understanding significantly contributes to improved health outcomes and a more satisfying healthcare experience. Administrators and managers grappling with the complexities of the pandemic derive immense benefits from explainable AI, facilitating resource allocation, policy development, and system optimization. Transparent AI models become invaluable tools for making data-driven decisions, ultimately optimizing resource utilization, and improving overall pandemic management strategies. Addressing challenges in AI frameworks in regard to Coronavirus identification, the discussion emphasizes the need for caution in model evaluation and the ongoing requirement for robust external validation. This underscores the pivotal importance of trustworthy models in the face of evolving healthcare scenarios.

For physicians, explainability emerges as a cornerstone in enhancing trust and confidence in AI-generated suggestions. The collaborative approach between healthcare professionals and AI technologies leads to more confident and informed clinical decisions, thereby improving the quality of care and optimizing treatment plans. Generative AI models, widely employed in education and research, are showcased for their impactful role in the scientific community's response to the pandemic. Researchers benefit from the transparency of AI models, gaining insights into COVID-19 features, uncovering new avenues of investigation, and ensuring the reliability of datasets. In the global impact of COVID-19, AI stands as a transformative force, yet challenges associated with black-box models necessitate a shift toward Explainable Artificial Intelligence (XAI). The legal and ethical dimensions underscore the need for transparency and responsible practices in deploying AI in healthcare. The multifaceted exploration of XAI, its categories, and its applications in medical research and practice underscores its capacity to improve decision process as well as collaboration. The nuanced understanding regarding explanation types within ML models, along with the diverse roles and expectations of stakeholders, further enriches the discussion, emphasizing the interdisciplinary nature of AI development. As we navigate the evolving landscape of AI and its intersection with healthcare, ongoing collaboration, research, and ethical considerations emerge as indispensable. Striking a equilibrium amidst cutting-edge progress and individual understanding remains a central theme. This discussion not only sets the stage for continued research and development in XAI methodologies but also aims to narrow the divide amid capabilities of machine intelligent models as well as an interpretability needed for ethical and trustworthy applications in critical domains like medicine. In doing so, we pave the way for a future where AI serves as a valuable ally in public health and beyond, fostering responsible deployment and improving outcomes for individuals and communities worldwide.

In parallel, the acknowledgment of the immense importance of explainability in generative AI models within the context of COVID-19 is crucial. In instances where XAI is not yet integrated into current generative AI models, a careful and considered approach by healthcare professionals and researchers becomes imperative. The absence of XAI capabilities may limit the understanding of internal decision-making processes, emphasizing the need for responsible practices. However, proactive measures can be implemented to ensure informed usage, emphasizing the collaborative efforts between healthcare professionals and AI developers. This collaboration not only advances the field but also enhances the credibility, trustworthiness, and ethical use of these tools in medical practice and research. The inclusion of mechanisms allowing for explainability provides insights into generative AI models' decision-making processes, increasing transparency and fostering trust in the generated outputs. This collective endeavor aligns with the overarching goal of responsible AI deployment, ensuring that technological advancements positively impact healthcare outcomes while upholding ethical standards and trust within the medical community and beyond.

References

Abumalloh, R. A., Asadi, S., Nilashi, M., Minaei-Bidgoli, B., Nayer, F. K., Samad, S., . . . Ibrahim, O. (2021). The impact of coronavirus pandemic (COVID-19) on education: The role of virtual and remote

An International Journal JSCDSS E-ISSN: 2289-8603

laboratories in education. *Technology in Society, 67*, 101728.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138-52160.

Ahmadi, H., Gholamzadeh, M., Shahmoradi, L., Nilashi, M., & Rashvand, P. (2018). Diseases diagnosis using fuzzy logic methods: A systematic and meta-analysis review. *Computer Methods and Programs in Biomedicine, 161*, 145-172.

Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database, 2020*, baaa010.

Ajunwa, I. (2020). An Auditing Imperative for Automated Hiring Systems. *Harv. JL & Tech., 34*, 621.

Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access, 6*, 14410-14430.

Akhtar, N., Mian, A., Kardan, N., & Shah, M. (2021). Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access, 9*, 155161-155196.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., . . . Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion, 99*, 101805.

Alikhademi, K., Richardson, B., Drobina, E., & Gilbert, J. E. (2021). Can explainable AI explain unfairness? A framework for evaluating explainable AI. *arXiv preprint arXiv:2106.07483*.

AlZubi, A. A., Alarifi, A., & Al-Maitah, M. (2020). Deep brain simulation wearable IoT sensor device based Parkinson brain disorder detection using heuristic tubu optimized sequence modular neural network. *Measurement, 161*, 107887.

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Consortium, P. Q. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making, 20*, 1-9.

Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences, 11*(11), 5088.

Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & society, 35*, 611-623.

Arora, P., Boyne, D., Slater, J. J., Gupta, A., Brenner, D. R., & Druzdzel, M. J. (2019). Bayesian networks for risk prediction using real-world data: a tool for precision medicine. *Value in Health, 22*(4), 439-445.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., . . . Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion, 58*, 82-115.

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: focus on clinicians. *Journal of medical Internet research, 22*(6), e15154.

Bagaric, M., Svilar, J., Bull, M., Hunter, D., & Stobbs, N. (2021). The Solution to the Pervasive Bias and Discrimination in the Criminal Justice: Transparent Artificial Intelligence. *American Criminal Law Review, 59*(1).

Balagurunathan, Y., Mitchell, R., & El Naqa, I. (2021). Requirements and reliability of AI in the medical context. *Physica Medica, 83*, 72-78.

Bashir, A. K., Victor, N., Bhattacharya, S., Huynh-The, T., Chengoden, R., Yenduri, G., . . . Liyanage, M. (2023). Federated Learning for the Healthcare Metaverse: Concepts, Applications, Challenges, and Future Directions. *IEEE Internet of Things Journal*.

Bellucci, M., Delestre, N., Malandain, N., & Zanni-Merk, C. (2021). Towards a terminology for a fully contextualized XAI. *Procedia Computer Science, 192*, 241-250.

Bhattacharya, A. (2022). *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more*: Packt Publishing Ltd.

Biran, O., & Cotton, C. (2017). *Explanation and justification in machine learning: A survey.* Paper presented at the IJCAI-17 workshop on explainable AI (XAI).

Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2023). Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 1-60.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and information technology, 15*, 209-227.

Burnham, J. P., Rojek, R. P., & Kollef, M. H. (2018). Catheter removal and outcomes of multidrug-resistant central-line-associated bloodstream infection. *Medicine, 97*(42).

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). *The role of explanations on trust and reliance in clinical decision support systems.* Paper presented at the 2015 international conference on healthcare informatics.

Callaway, E. (2023). The next generation of coronavirus vaccines. *Nature, 614*, 22-25.

Canny, A., Donaghy, E., Murray, V., Campbell, L., Stonham, C., Bush, A., . . . Daines, L. (2023). Patient views on asthma diagnosis and how a clinical decision support system could help: A

An International Journal JSCDSS E-ISSN: 2289-8603

qualitative study. *Health Expectations, 26*(1), 307-317.

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI & society, 35*(2), 309-317.

Carter, A. (2018). Cathy O'Neil (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, New York, St. Martin's Press and Virginia Eubanks (2018) Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, New York, Broadway Books.

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., . . . Rao, R. M. (2017). *Interpretability of deep learning models: A survey of results.* Paper presented at the 2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI).

D'Agostino Sr, R. B., Pencina, M. J., Massaro, J. M., & Coady, S. (2013). Cardiovascular disease risk assessment: insights from Framingham. *Global heart, 8*(1), 11-23.

Dalenberg, D. J. (2018). Preventing discrimination in the automated targeting of job advertisements. *Computer law & security review, 34*(3), 615-627.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Delahanty, R. J., Alvarez, J., Flynn, L. M., Sherwin, R. L., & Jones, S. S. (2019). Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. *Annals of emergency medicine, 73*(4), 334-344.

Ding, W., Abdel-Basset, M., Hawash, H., & Ali, A. M. (2022). Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances, 4*(1), eaao5580.

Duque Anton, S. D., Schneider, D., & Schotten, H. D. (2022). *On Explainability in AI-Solutions: A Cross-Domain Survey.* Paper presented at the International Conference on Computer Safety, Reliability, and Security.

Emmert-Streib, F., Yli-Harja, O., & Dehmer, M. (2020). Explainable artificial intelligence and machine learning: A reality rooted perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(6), e1368.

Erasmus, A., Brunet, T. D., & Fisher, E. (2021). What is interpretability? *Philosophy & Technology, 34*(4), 833-862.

Fabre, V., Amoah, J., Cosgrove, S. E., & Tamma, P. D. (2019). Antibiotic therapy for Pseudomonas aeruginosa bloodstream infections: how long is long enough? *Clinical Infectious Diseases, 69*(11), 2011-2014.

Fuhrman, J. D., Gorre, N., Hu, Q., Li, H., El Naqa, I., & Giger, M. L. (2022). A review of explainable and interpretable AI with applications in COVID-19 imaging. *Medical Physics, 49*(1), 1-14.

Gaviglio, A. M., Skinner, M. W., Lou, L. J., Finkel, R. S., Augustine, E. F., & Goldenberg, A. J. (2023). *Gene-targeted therapies: Towards equitable development, diagnosis, and access.* Paper presented at the American Journal of Medical Genetics Part C: Seminars in Medical Genetics.

Ghane, M., Ang, M. C., Nilashi, M., & Sorooshian, S. (2022). Enhanced decision tree induction using evolutionary techniques for Parkinson's disease classification. *Biocybernetics and Biomedical Engineering, 42*(3), 902-920.

Ghoshal, B., & Tucker, A. (2020). Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. *arXiv preprint arXiv:2003.10769*.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). *Explaining explanations: An overview of interpretability of machine learning.* Paper presented at the 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA).

Gkontra, P., Quaglio, G., Garmendia, A. T., & Lekadir, K. (2023). Challenges of Machine Learning and AI (What Is Next?), Responsible and Ethical AI *Clinical Applications of Artificial Intelligence in Real-World Data* (pp. 263-285): Springer.

Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J. P., Yordanova, K., Vered, M., . . . Pulignano, V. (2023). A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artificial intelligence review, 56*(4), 3473-3504.

Griffin, T. A., Green, B. P., & Welie, J. V. (2023). The ethical agency of AI developers. *AI and Ethics*, 1-10.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR), 51*(5), 1-42.

Guillemé, M., Masson, V., Rozé, L., & Termier, A. (2019). *Agnostic local explanation for time series classification.* Paper presented at the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI).

Habuza, T., Navaz, A. N., Hashim, F., Alnajjar, F., Zaki, N., Serhani, M. A., & Statsenko, Y. (2021). AI applications in robotics, diagnostic image analysis and precision medicine: current limitations, future trends, guidelines on CAD systems for medicine. *Informatics in Medicine Unlocked, 24*, 100596.

Hacker, P., Engel, A., & Mauer, M. (2023). *Regulating ChatGPT and other large generative AI models.* Paper presented at the Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.

Hagras, H. (2018). Toward human-understandable, explainable AI. *Computer, 51*(9), 28-36.

Hall, K., Chang, V., & Jayne, C. (2022). A review on Natural Language Processing Models for COVID-19 research. *Healthcare Analytics*, 100078.

Harris, P. N., Tambyah, P. A., Lye, D. C., Mo, Y., Lee, T. H., Yilmaz, M., . . . Bassetti, M. (2018). Effect of piperacillin-tazobactam vs meropenem on 30-day mortality for patients with E coli or Klebsiella pneumoniae bloodstream infection and ceftriaxone resistance: a randomized clinical trial. *Jama, 320*(10), 984-994.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., . . . Hussain, A. (2023). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 1-30.

Hicks, S. A., Eskeland, S., Lux, M., de Lange, T., Randel, K. R., Jeppsson, M., . . . Riegler, M. (2018). *Mimir: an automatic reporting and reasoning system for deep learning based analysis in the medical domain.* Paper presented at the Proceedings of the 9th ACM Multimedia Systems Conference.

Howard, J. J., Sirotin, Y. B., Tipton, J. L., & Vemury, A. R. (2021). Reliability and validity of image-based and self-reported skin phenotype metrics. *IEEE Transactions on Biometrics, Behavior, and Identity Science, 3*(4), 550-560.

Hu, S., Gao, Y., Niu, Z., Jiang, Y., Li, L., Xiao, X., . . . Xia, J. (2020). Weakly supervised deep learning for covid-19 infection detection and classification from ct images. *IEEE Access, 8*, 118869-118883.

Husky, M. M., Kovess-Masfety, V., & Swendsen, J. D. (2020). Stress and anxiety among university students in France during Covid-19 mandatory confinement. *Comprehensive psychiatry, 102*, 152191.

Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks, 3*, 58-73.

Jiang, J., Kahai, S., & Yang, M. (2022). Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *International Journal of Human-Computer Studies, 165*, 102839.

Kam, H. J., & Kim, H. Y. (2017). Learning representations for the early detection of sepsis with deep neural networks. *Computers in Biology and Medicine, 89*, 248-255.

Karim, M. R., Döhmen, T., Rebholz-Schuhmann, D., Decker, S., Cochez, M., & Beyan, O. (2020). DeepCOVIDExplainer: Explainable COVID-19 diagnosis based on chest X-ray images. *arXiv preprint arXiv:2004.04582*.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine, 17*, 1-9.

Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society, 22*(14), 2081-2096.

Kenny, E. M., Ford, C., Quinn, M., & Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence, 294*, 103459.

Kök, İ., Okay, F. Y., Muyanlı, Ö., & Özdemir, S. (2023). Explainable artificial intelligence (xai) for internet of things: a survey. *IEEE Internet of Things Journal*.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., . . . Maningo, J. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health, 2*(2), e0000198.

Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L., & Bonham, V. L. (2018). Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Affairs, 37*(5), 780-785.

Lane, H. C., Core, M. G., Van Lent, M., Solomon, S., & Gomboc, D. (2005). *Explainable Artificial Intelligence for Training and Tutoring.* Paper presented at the AIED.

Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior, 123*, 106878.

Lee, H.-C., Yoon, S. B., Yang, S.-M., Kim, W. H., Ryu, H.-G., Jung, C.-W., . . . Lee, K. H. (2018). Prediction of acute kidney injury after liver transplantation: machine learning approaches vs. logistic regression model. *Journal of clinical medicine, 7*(11), 428.

Li, M. D., Arun, N. T., Gidwani, M., Chang, K., Deng, F., Little, B. P., . . . O'Shea, A. (2020). Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. *Radiology: Artificial Intelligence, 2*(4), e200079.

Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., . . . Dou, D. (2022). Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems, 64*(12), 3197-3234.

An International Journal JSCDSS E-ISSN: 2289-8603

Lisboa, P., Saralajew, S., Vellido, A., Fernández-Domenech, R., & Villmann, T. (2023). The coming of age of interpretable and explainable machine learning models. *Neurocomputing, 535*, 25-39.

Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, 107161.

Lopes, P., Silva, E., Braga, C., Oliveira, T., & Rosado, L. (2022). XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. *Applied Sciences, 12*(19), 9423.

Maguolo, G., & Nanni, L. (2021). A critic evaluation of methods for COVID-19 automatic detection from X-ray images. *Information fusion, 76*, 1-7.

Mahya, P., & Fürnkranz, J. (2023). An Empirical Comparison of Interpretable Models to Post-Hoc Explanations. *AI, 4*(2), 426-436.

McNally, J. G., Karpova, T., Cooper, J., & Conchello, J. A. (1999). Three-dimensional imaging by deconvolution microscopy. *Methods, 19*(3), 373-385.

Mehta, O., Liao, Z., Jenkinson, M., Carneiro, G., & Verjans, J. (2022). Machine learning in medical imaging–clinical applications and challenges in computer vision. *Artificial Intelligence in Medicine: Applications, Limitations and Future Directions*, 79-99.

Melis, D. A., Kaur, H., Daumé III, H., Wallach, H., & Vaughan, J. W. (2021). *From human explanation to model interpretability: A framework based on weight of evidence.* Paper presented at the Proceedings of the AAAI Conference on Human Computation and Crowdsourcing.

Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ digital medicine, 6*(1), 120.

Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., . . . Duffy, N. (2024). Evolving deep neural networks *Artificial intelligence in the age of neural networks and brain computing* (pp. 269-287): Elsevier.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). *Interpretable machine learning–a brief history, state-of-the-art and challenges.* Paper presented at the Joint European conference on machine learning and knowledge discovery in databases.

Muniz, A., Liu, H., Lyons, K., Pahwa, R., Liu, W., Nobre, F., & Nadal, J. (2010). Comparison among probabilistic neural network, support vector machine and logistic regression for evaluating the effect of subthalamic stimulation in Parkinson disease on ground reaction force during gait. *Journal of biomechanics, 43*(4), 720-726.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies, 169*, 102941.

Natesan Ramamurthy, K., Vinzamuri, B., Zhang, Y., & Dhurandhar, A. (2020). Model agnostic multilevel explanations. *Advances in neural information processing systems, 33*, 5968-5979.

Nilashi, M., Abumalloh, R. A., Alghamdi, A., Minaei-Bidgoli, B., Alsulami, A. A., Thanoon, M., . . . Samad, S. (2021). What is the impact of service quality on customers' satisfaction during COVID-19 outbreak? New findings from online reviews analysis. *Telematics and Informatics, 64*, 101693.

Nilashi, M., Abumalloh, R. A., Almulihi, A., Alrizq, M., Alghamdi, A., Ismail, M. Y., . . . Asadi, S. (2023). Big social data analysis for impact of food quality on travelers' satisfaction in eco-friendly hotels. *ICT Express, 9*(2), 182-188.

Nilashi, M., Abumalloh, R. A., Alrizq, M., Alghamdi, A., Samad, S., Almulihi, A., . . . Mohd, S. (2022). What is the impact of eWOM in social network sites on travel decision-making during the COVID-19 outbreak? A two-stage methodology. *Telematics and Informatics, 69*, 101795.

Nilashi, M., Abumalloh, R. A., Alrizq, M., Almulihi, A., Alghamdi, O., Farooque, M., . . . Ahmadi, H. (2022). A hybrid method to solve data sparsity in travel recommendation agents using fuzzy logic approach. *Mathematical Problems in Engineering, 2022*.

Nilashi, M., Abumalloh, R. A., Minaei-Bidgoli, B., Samad, S., Yousoof Ismail, M., Alhargan, A., & Abdu Zogaan, W. (2022). Predicting parkinson's disease progression: Evaluation of ensemble methods in machine learning. *Journal of healthcare engineering, 2022*.

Nilashi, M., Abumalloh, R. A., Minaei-Bidgoli, B., Zogaan, W. A., Alhargan, A., Mohd, S., . . . Samad, S. (2022). Revealing travellers' satisfaction during COVID-19 outbreak: moderating role of service quality. *Journal of Retailing and Consumer Services, 64*, 102783.

Nilashi, M., Abumalloh, R. A., Mohd, S., Azhar, S. N. F. S., Samad, S., Thi, H. H., . . . Alghamdi, A. (2023). COVID-19 and sustainable development goals: A bibliometric analysis and SWOT analysis in Malaysian context. *Telematics and Informatics, 76*, 101923.

Nilashi, M., Abumalloh, R. A., Yusuf, S. Y. M., Thi, H. H., Alsulami, M., Abosaq, H., . . . Alghamdi, A. (2023). Early diagnosis of Parkinson's disease: A combined method using deep learning and neuro-fuzzy techniques. *Computational biology and chemistry, 102*, 107788.

Nilashi, M., Ahmadi, H., Shahmoradi, L., Ibrahim, O., & Akbari, E. (2019). A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique. *Journal of infection and public health, 12*(1), 13-20.

An International Journal JSCDSS E-ISSN: 2289-8603

Nilashi, M., Ahmadi, H., Sheikhtaheri, A., Naemi, R., Alotaibi, R., Alarood, A. A., . . . Zhao, J. (2020). Remote tracking of Parkinson's disease progression using ensembles of deep belief network and self-organizing map. *Expert Systems with Applications*, *159*, 113562.

Nilashi, M., Asadi, S., Minaei-Bidgoli, B., Abumalloh, R. A., Samad, S., Ghabban, F., & Ahani, A. (2021). Recommendation agents and information sharing through social media for coronavirus outbreak. *Telematics and Informatics*, *61*, 101597.

Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering*, *106*, 212-223.

Nilashi, M., Bin Ibrahim, O., Mardani, A., Ahani, A., & Jusoh, A. (2018). A soft computing approach for diabetes disease classification. *Health Informatics Journal*, *24*(4), 379-393.

Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L., & Farahmand, M. (2018). A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques. *Biocybernetics and Biomedical Engineering*, *38*(1), 1-15.

Nilashi, M., Ibrahim, O., Samad, S., Ahmadi, H., Shahmoradi, L., & Akbari, E. (2019). An analytical method for measuring the Parkinson's disease progression: A case on a Parkinson's telemonitoring dataset. *Measurement*, *136*, 545-557.

Nilashi, M., Rupani, P. F., Rupani, M. M., Kamyab, H., Shao, W., Ahmadi, H., . . . Aljojo, N. (2019). Measuring sustainability through ecological sustainability and human sustainability: A machine learning approach. *Journal of Cleaner Production*, *240*, 118162.

Nilashi, M., Samad, S., Ahani, A., Ahmadi, H., Alsolami, E., Mahmoud, M., . . . Alarood, A. A. (2021). Travellers decision making through preferences learning: A case on Malaysian spa hotels in TripAdvisor. *Computers & Industrial Engineering*, *158*, 107348.

Ozyurt, F., Tuncer, T., & Subasi, A. (2021). An automated COVID-19 detection based on fused dynamic exemplar pyramid feature extraction and hybrid feature selection using deep learning. *Computers in Biology and Medicine*, *132*, 104356.

Packin, N. G., & Lev Aretz, Y. (2018). Learning algorithms and discrimination. *Book Chapter: Learning Algorithms and Discrimination In RESEARCH HANDBOOK OF ARTIFICIAL INTELLIGENCE AND LAW (Woodrow Barfield & Ugo Pagallo eds.)(2018 Forthcoming), Baruch College Zicklin School of Business Research Paper*(2018-04), 03.

Park, H. (2022). Providing post-hoc explanation for node representation learning models through inductive conformal predictions. *IEEE Access*, *11*, 1202-1212.

Pedrero-Sánchez, J. F., Belda-Lois, J.-M., Serra-Ano, P., Ingles, M., & Lopez-Pascual, J. (2022). Classification of healthy, Alzheimer and Parkinson populations with a multi-branch neural network. *Biomedical Signal Processing and Control*, *75*, 103617.

Peres, R., Schreier, M., Schweidel, D., & Sorescu, A. (2023). On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing*.

Picardi, C., Hawkins, R., Paterson, C., & Habli, I. (2019). *A pattern for arguing the assurance of machine learning in medical diagnosis systems.* Paper presented at the Computer Safety, Reliability, and Security: 38th International Conference, SAFECOMP 2019, Turku, Finland, September 11–13, 2019, Proceedings 38.

Qureshi, R., Irfan, M., Ali, H., Khan, A., Nittala, A. S., Ali, S., . . . Shah, Z. (2023). Artificial Intelligence and Biosensors in Healthcare and its Clinical Relevance: A Review. *IEEE Access*.

Rahmani, A. M., Yousefpoor, E., Yousefpoor, M. S., Mehmood, Z., Haider, A., Hosseinzadeh, M., & Ali Naqvi, R. (2021). Machine learning (ML) in medicine: Review, applications, and challenges. *Mathematics*, *9*(22), 2970.

Rao, S., Böhle, M., & Schiele, B. (2022). *Towards better understanding attribution methods.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Ren, K., Zheng, T., Qin, Z., & Liu, X. (2020). Adversarial attacks and defenses in deep learning. *Engineering*, *6*(3), 346-360.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *" Why should i trust you?" Explaining the predictions of any classifier.* Paper presented at the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). *Anchors: High-precision model-agnostic explanations.* Paper presented at the Proceedings of the AAAI conference on artificial intelligence.

Rupani, P. F., Nilashi, M., Abumalloh, R. e., Asadi, S., Samad, S., & Wang, S. (2020). Coronavirus pandemic (COVID-19) and its natural environmental impacts. *International Journal of Environmental Science and Technology*, *17*, 4655-4666.

Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, *263*, 110273.

An International Journal JSCDSS E-ISSN: 2289-8603

Samek, W., & Müller, K.-R. (2019). Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, 5-22.

Samih, A., Adadi, A., & Berrada, M. (2019). *Towards a knowledge based explainable recommender systems.* Paper presented at the Proceedings of the 4th International Conference on Big Data and Internet of Things.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-cam: Visual explanations from deep networks via gradient-based localization.* Paper presented at the Proceedings of the IEEE international conference on computer vision.

Sharma, R., Gulati, A., & Chopra, K. (2023). Artificial Intelligence (AI) and Machine Learning (ML): An Innovative Cross-Talk Perspective and Their Role in the Healthcare Industry *Artificial Intelligence and Machine Learning in Healthcare* (pp. 9-38): Springer.

Sheu, R.-K., & Pardeshi, M. S. (2022). A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. *Sensors, 22*(20), 8068.

Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K., & Weisz, J. D. (2022). *Investigating explainability of generative AI for code through scenario-based design.* Paper presented at the 27th International Conference on Intelligent User Interfaces.

Taheri, S., Asadi, S., Nilashi, M., Abumalloh, R. A., Ghabban, N. M., Yusuf, S. Y. M., . . . Samad, S. (2021). A literature review on beneficial role of vitamins and trace elements: Evidence from published clinical studies. *Journal of Trace Elements in Medicine and Biology, 67*, 126789.

Taresh, M. M., Zhu, N., Ali, T. A. A., Hameed, A. S., & Mutar, M. L. (2021). Transfer learning to detect covid-19 automatically from x-ray images using convolutional neural networks. *International Journal of Biomedical Imaging, 2021*, 1-9.

Thakur, S., & Kumar, A. (2021). X-ray and CT-scan-based automated detection and classification of covid-19 using convolutional neural networks (CNN). *Biomedical Signal Processing and Control, 69*, 102920.

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems, 32*(11), 4793-4813.

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*.

Tong, L., Shi, W., Isgut, M., Zhong, Y., Lais, P., Gloster, L., . . . Wang, M. D. (2023). Integrating Multi-omics Data with EHR for Precision Medicine Using Advanced Artificial Intelligence. *IEEE Reviews in Biomedical Engineering*.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine, 25*(1), 44-56.

Tosun, A. B., Pullara, F., Becich, M. J., Taylor, D. L., Fine, J. L., & Chennubhotla, S. C. (2020). Explainable AI (xAI) for anatomic pathology. *Advances in Anatomic Pathology, 27*(4), 241-250.

Tucci, V., Saary, J., & Doyle, T. E. (2022). Factors influencing trust in medical artificial intelligence for healthcare professionals: A narrative review. *Journal of Medical Artificial Intelligence, 5*.

Tuncer, T., Ozyurt, F., Dogan, S., & Subasi, A. (2021). A novel Covid-19 and pneumonia classification method based on F-transform. *Chemometrics and intelligent laboratory systems, 210*, 104256.

Udeshi, S., Peng, S., Woo, G., Loh, L., Rawshan, L., & Chattopadhyay, S. (2022). Model agnostic defence against backdoor attacks in machine learning. *IEEE Transactions on Reliability, 71*(2), 880-895.

Vale, D., El-Sharif, A., & Ali, M. (2022). Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI and Ethics, 2*(4), 815-826.

Veiber, L., Allix, K., Arslan, Y., Bissyandé, T. F., & Klein, J. (2020). *Challenges Towards {Production-Ready} Explainable Machine Learning.* Paper presented at the 2020 USENIX Conference on Operational Machine Learning (OpML 20).

Vilone, G., & Longo, L. (2021). A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods. *Frontiers in artificial intelligence, 4*, 717899.

Vorm, E., & Combs, D. J. (2022). Integrating transparency, trust, and acceptance: The intelligent systems technology acceptance model (ISTAM). *International Journal of Human–Computer Interaction, 38*(18-20), 1828-1845.

Wang, L., Lin, Z. Q., & Wong, A. (2020). Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports, 10*(1), 19549.

WHO. (2023). WHO Coronavirus (COVID-19) Dashboard. Retrieved from https://covid19.who.int/

Wu, J., Peck, D., Hsieh, S., Dialani, V., Lehman, C. D., Zhou, B., . . . Patterson, G. (2018). *Expert identification of visual primitives used by CNNs during mammogram classification.* Paper presented at the Medical Imaging 2018: Computer-Aided Diagnosis.

Xie, Y., Gao, G., & Chen, X. A. (2019). Outlining the design space of explainable intelligent systems for medical diagnosis. *arXiv preprint arXiv:1902.06019*.

Yang, C. C. (2022). Explainable artificial intelligence for predictive modeling in healthcare. *Journal of healthcare informatics research, 6*(2), 228-239.

Ye, Q., Xia, J., & Yang, G. (2021). *Explainable AI for COVID-19 CT classifiers: an initial comparison*

An International Journal JSCDSS E-ISSN: 2289-8603

study. Paper presented at the 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS).

Zhang, J., Yu, L., Chen, D., Pan, W., Shi, C., Niu, Y., . . . Cheng, Y. (2021). Dense GAN and multi-layer attention based lesion segmentation method for COVID-19 CT images. *Biomedical Signal Processing and Control, 69*, 102901.

Zhang, Y. (2017). Can a smartphone diagnose parkinson disease? a deep neural network method and telediagnosis system implementation. *Parkinson's disease, 2017*.

Zheng, Q., Delingette, H., & Ayache, N. (2019). Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow. *Medical image analysis, 56*, 80-95.

An International Journal
JSCDSS
E-ISSN: 2289-8603