Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 9, 2025

UNDERSTANDING MACHINE LEARNING–DRIVEN PREDICTIVE MODELS IN HEALTHCARE

Muhammad Hammad u Salam[*1], Shujaat Ali Rathore[2], Muhammad Irfan[3]

[*1,2]*Department of Computer Science & Information Technology, University of Kotli, Azad Jammu and Kashmir.*
[3]*Department of Computer Science, NCBA&E, Sub-Campus Multan, 60000, Pakistan*

hammad.salam@uokajk.edu.pk

DOI: https://doi.org/10.5281/zenodo.17164999

Keywords
Interpretability, machine learning, model-agnostic, model-specific, prediction models, healthcare AI, explainable AI (XAI), clinical decision support, transparency, accountability, patient safety

Article History
Received: 11 June 2025
Accepted: 21 August 2025
Published: 20 September 2025

Corresponding Author: * Muhammad Hammad u Salam

Abstract

In high-stakes fields like healthcare, it's essential for machine learning (ML) models to be interpretable, meaning their predictions are easily understood and explained to end-users. This transparency allows healthcare professionals to make informed, data-driven decisions, leading to more personalized care and a higher quality of service. Interpretability methods can be broadly categorized into two groups. The first group provides local interpretability, focusing on explaining individual predictions, while the second group offers global interpretability, summarizing the model's behavior across an entire population. Another way to classify these approaches is by their dependency on the model itself. Model-specific techniques are tailored for a particular type of model, like a neural network, whereas model-agnostic methods can be used to interpret predictions from any ML model. This overview explores various interpretability approaches for structured data and provides practical examples of their application in healthcare. We'll discuss how these methods can be used to predict health outcomes, optimize treatment plans, and improve the efficiency of screening for specific conditions. Finally, we'll outline future directions for interpretable ML, emphasizing the need for new algorithmic solutions that can facilitate reliable, ML-driven decision-making in critical healthcare scenarios.

INTRODUCTION

Recent advancements in technology and the digital revolution have led to widespread adoption of artificial intelligence (AI) across numerous sectors, including healthcare (Adadi & Berrada, 2018). Among these advancements, machine learning (ML) has demonstrated exceptional capability in identifying and analyzing complex data patterns (Murdoch et al., 2018). In particular, ML-based models are now widely used for structured data analysis, often surpassing traditional statistical approaches in terms of predictive performance. However, despite their success, many of these sophisticated ML systems function as "black boxes," offering little to no insight into how predictions are generated. This lack of transparency, intuitiveness, and interpretability presents significant challenges—especially in critical decision-making environments like healthcare, where understanding the basis of predictions is essential. To be useful and trustworthy in clinical settings, ML models must be both understandable and explainable. Understandability refers to how easily the end-user can grasp the rationale behind a prediction, while interpretability and explainability—terms often used interchangeably (Carvalho et al.,

2019)—focus on the extent to which a human can comprehend the internal decision-making process of the model. As Adadi and Berrada (2018) describe, an interpretable system is one where its operations can be understood by humans. In healthcare, this is particularly important due to the high-stakes nature of the field. Over the past few years, there has been growing interest in ML-powered healthcare applications (Ahmad, Eckert, & Teredesai, 2018), such as predicting patient disease risks, likelihood of readmission, or the need for medical intervention.For healthcare professionals to trust and adopt these systems, they must be able to understand and assess the reasoning behind the model's outputs. This growing demand for transparency has driven the need for interpretable machine learning techniques (Carvalho et al., 2019). A major limitation to widespread adoption of ML in healthcare lies in the complexity of these models, which often require a high level of technical knowledge and may only be validated for narrowly defined tasks or datasets. For example, Obermeyer et al. (2019) found evidence of racial bias in U.S. healthcare algorithms, highlighting how lack of interpretability can compromise trust and fairness. Furthermore, models built on homogeneous datasets may not generalize well to diverse patient populations, and many prioritize prediction accuracy over transparency (Ahmad et al., 2018; Elshawi et al., 2019). In clinical practice, accuracy is essential—but so is understanding why a prediction was made.The field of interpretable ML has therefore emerged to develop models that are both accurate and transparent, or to adapt complex models into more understandable forms. This includes transforming opaque black-box models into white-box models with interpretable predictions. However, interpretability itself is a complex, evolving concept (Gilpin et al., 2019; Hall & Gill, 2018), with many overlapping terms such as model comprehensibility, locality, or mental fit (Breiman, 2001; Piltaver et al., 2016; Bibal & Frenay, 2016). Generally, interpretability refers to how well a human can understand the cause of a model's decision (Miller, 2019), and it plays a central role in ensuring transparency in AI systems.There is no universally agreed-upon definition or metric for interpretability, as it is often subjective and context-dependent. Still, an interpretable model is typically one that balances accuracy, transparency, and efficiency—particularly the time it takes for a user to understand it (Bibal & Frenay, 2016). The concept of "explainable AI" was first introduced by van Lent et al. (2004) to describe systems that clearly articulate their behavior, and it has since become an essential goal in healthcare AI applications. Interpretability is now considered a fundamental requirement in the development of trustworthy AI systems (Carvalho et al., 2019), driven by motivations such as safety, fairness, regulation, and societal impact.In healthcare, an interpretable model should clearly communicate its predictions to the end-user, allowing clinicians to accept or reject its recommendations based on sound reasoning (Ahmad et al., 2018). Simple models like decision trees or logistic regression are inherently interpretable, while more complex models—such as ensemble classifiers or deep neural networks—often require post hoc explanation techniques to make their predictions understandable. In response to ethical and legal concerns, especially within Europe, the General Data Protection Regulation (GDPR) mandates that decisions made by automated systems be explainable (Wachter et al., 2017; Greene et al., 2019; Wallace & Castro, 2018).Despite this, some researchers question whether complete interpretability should be required. For example, AI pioneer Geoffrey Hinton argues that people often cannot explain their own reasoning processes and therefore should not expect machines to do so (Wang & Preininger, 2019). On the other hand, Rudin (2019) advocates for the development of inherently interpretable models rather than attempting to explain black-box ones, warning that reliance on post hoc explanations can lead to poor practices and potentially harmful outcomes. In her view, interpretable alternatives—such as sparse scoring systems (Ustun & Rudin, 2017)—are particularly suitable for high-stakes domains like healthcare.This ongoing debate continues to be explored in recent research (Jia et al., 2020), where experts express differing views on the balance between accuracy and interpretability. In clinical environments where healthcare workers face heavy workloads, large datasets, and complex decision-making processes, interpretable ML tools can significantly enhance decision support and improve patient outcomes

(Ahmad et al., 2018; Carvalho et al., 2019). ML is especially effective in identifying disease risks and optimizing resource allocation, but high-performing models are often difficult to interpret (e.g., deep learning networks; Michalopoulos et al., 2020). Improving interpretability is thus essential for practical adoption, and should ideally involve collaboration between data scientists and clinical experts (Vellido, 2019).

It's also important to recognize the limitations of ML models, such as dataset shifts, unintentional biases, overfitting to confounders, and difficulties in generalization (Kelly et al., 2019). While deep learning has driven major advances in AI, its models remain largely uninterpretable. Some researchers argue that AI can be safe and effective even without complete transparency, as long as human oversight is maintained (Ren, 2020). Others believe that interpretability is crucial to responsible AI deployment. Notably, patients themselves often value model effectiveness over interpretability (Jia et al., 2020), raising important questions about how and when interpretability should be prioritized.

This article introduces a taxonomy of interpretability techniques for ML-based predictive models in healthcare. Its purpose is to offer clear, accessible definitions for readers new to the field and to support healthcare professionals in understanding the foundational concepts of explainable ML. It is important to note that this discussion focuses on structured data applications, and does not cover interpretability methods related to unstructured data such as medical imaging, clinical text, or biosignals. For a broader exploration of explainable AI in healthcare, readers are encouraged to consult reviews by Holzinger et al. (2019) and London (2019). Finally, recent work on responsible AI in healthcare by Wiens et al. (2019) proposes a framework for developing trustworthy ML systems through interdisciplinary collaboration, aligning model transparency with policy, ethics, and real-world impact.

CATEGORIZATION OF ML MODELS INTERPRETABILITY

The interpretability of machine learning (ML) models can be categorized using various frameworks, such as intrinsic versus post hoc methods, and also based on the phase in which interpretability techniques are applied—premodel, in-model, or postmodel (Carvalho et al., 2019; Molnar, 2020).

Intrinsic interpretability

Refers to models that are inherently transparent due to their simplicity, such as linear regression or decision trees. These models allow a basic understanding of the decision-making process without the need for additional interpretability tools. However, even these models have limitations, particularly in their ability to capture complex nonlinear relationships, which are common in healthcare data.

In contrast, **post hoc (or postmodel) interpretability** involves applying interpretability methods after the model has been trained. These methods are typically used with more complex, black-box models that do not provide direct insight into their internal workings (Molnar, 2020).

Another way to classify interpretability methods is by the point at which they are applied in the ML pipeline:

- **Premodel interpretability** refers to techniques used before model selection or training. These include exploratory data analysis methods like descriptive statistics, data visualization, Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and clustering. While these approaches can help detect patterns or relationships in the data, they often lack direct clinical interpretability—particularly when using transformations like PCA or embeddings that obscure the original features' meanings.
- **In-model interpretability** is characteristic of models that are designed to be interpretable from the outset. These models integrate transparency into their structure, allowing users to understand how predictions are made without the need for external explanation tools.
- **Postmodel interpretability**, also known as post hoc interpretation, is applied after model development. These methods aim to shed light on how a trained model functions without modifying the model itself. As Murdoch et al. (2018) describe, they

are useful for extracting insights from complex models while maintaining their original architecture.

Interpretability can also be categorized based on the **type of output** or **insights provided**. Molnar (2020) outlines several types of interpretability methods:

- **Feature summary statistics** provide aggregated insights into the importance or influence of individual features on model predictions.
- **Feature summary visualizations** transform statistical outputs into graphical representations, making patterns and influences easier to interpret.
- **Model internals interpretation** involves directly examining the inner workings of the model—such as coefficients in linear models or decision rules in trees—which is only feasible with inherently interpretable models.
- **Data point–level explanations** focus on interpreting individual predictions, often by identifying similar instances or through local approximation methods. While effective in domains like image or text analysis, these are less suited to high-dimensional, tabular healthcare data.
- **Surrogate models** attempt to approximate black-box models using simpler, interpretable models for explanation purposes.

A further distinction is made between **global** and **local interpretability**.

- **Global interpretability** involves understanding how the entire model behaves across all inputs and predictions, offering a high-level overview of the learned relationships (Bratko, 1997; Martens et al., 2008).
- **Local interpretability**, on the other hand, focuses on individual predictions or small regions of the input space, helping users understand why a specific decision was made (Hall et al., 2019; Stiglic et al., 2006). Both approaches are valuable in clinical practice and continue to evolve alongside advances in ML methodology.

In healthcare, the goals of ML evaluation differ from many other domains. While general ML applications may emphasize overall performance metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC), clinical applications often prioritize metrics like **sensitivity**, **specificity**, or **positive predictive value**, which have greater practical implications for patient care (Simon et al., 2019; Steyerberg, 2019). These performance considerations can also influence how interpretable a model needs to be in a clinical setting.

To support both experts and non-experts, we propose a straightforward framework for classifying ML interpretability approaches into two **non-mutually exclusive** categories:

1. **Model-specific vs. Model-agnostic**:
   - *Model-specific methods* are tailored to the internal structure of a particular model type.
   - *Model-agnostic methods* can be applied to any model, regardless of its design.
2. **Global vs. Local Interpretability**:
   - *Global methods* explain the model as a whole.
   - *Local methods* focus on explaining individual predictions.

This categorization helps make interpretability more accessible to a broad audience, including clinicians, researchers, and policy makers working in healthcare AI.

MODEL-SPECIFIC OR MODEL-AGNOSTIC INTERPRETABILITY

Interpretability techniques in machine learning (ML) can generally be divided into two main categories: **model-specific** and **model-agnostic**. These approaches are designed to help understand how predictions are made—either by directly analyzing the internal structure of the model or by using external tools to approximate the model's behavior.

Model-specific interpretability

Methods are tailored to particular types of ML models. For example, with decision trees, interpretability is often achieved by extracting the full set of decision rules used to arrive at a prediction. These techniques rely on accessing and analyzing the model's internal parameters—such as weights, splits, or feature importance—and are therefore limited to models with transparent architectures (Du, Liu, & Hu, 2019).

In contrast, **model-agnostic interpretability** treats the ML model as a black box and does not depend on the model's internal components. Instead, these

techniques typically operate **after the model has been trained (post hoc)** and analyze only the relationship between inputs and outputs (Molnar, 2020). This flexibility allows model-agnostic approaches to be applied to any type of ML model, including highly complex or opaque systems like deep neural networks (Du et al., 2019).

One common technique in model-agnostic interpretability involves building a **surrogate model**—a simpler, interpretable model such as a decision tree or linear regression—that mimics the behavior of the more complex model within a local decision boundary. This strategy provides a faithful approximation of the original model's predictions in specific regions, making it easier to understand how predictions are derived.

A foundational example of this approach was introduced by Craven and Shavlik (1994), who demonstrated how to extract interpretable rules from trained neural networks. Similarly, Bucila, Caruana, and Niculescu-Mizil (2006) introduced the concept of **model compression**, where a compact, interpretable model learns to reproduce the outputs of a larger, more complex one. This idea was later expanded and popularized under the term **knowledge distillation** (Hinton, Vinyals, & Dean, 2015), where a simple model is trained to imitate the performance of a deep network while retaining interpretability.

Another recent development in interpretability, applicable to both model-specific and model-agnostic frameworks, is the **GNNExplainer** (Ying et al., 2019). Designed for **Graph Neural Networks (GNNs)**—which are increasingly used for modeling complex relational data—GNNExplainer aims to uncover and visualize the most relevant components of graph data that influence the model's predictions. This tool provides both **local and global interpretability**, making it a valuable asset for analyzing the behavior of GNN-based systems.

Although a full discussion of GNNExplainer and related techniques is beyond the scope of this article, readers interested in these advanced methods can refer to foundational works by Hamilton et al. (2017). The continued development of tools like GNNExplainer is likely to play a key role in the broader adoption of interpretable ML in domains requiring graph-based data representation. As these methods evolve, they offer promising avenues for both improving transparency and diagnosing errors in complex models such as GNNs (Ying et al., 2019).

LOCAL OR GLOBAL INTERPRETABILITY

Interpretability in machine learning can be broadly categorized into **local** and **global** approaches, depending on the scope of explanation they provide (Molnar, 2020).

Local interpretability focuses on explaining the reasoning behind an individual prediction. This type of interpretability is achieved either by designing model architectures that inherently support explanations or by comparing the target data point to similar cases. For instance, in a healthcare context, one might highlight specific patient attributes that closely resemble a subset of other patients but differ from the general population. This provides insight into why a particular decision was made for that individual.

Although local interpretability was less emphasized in earlier research, recent years have seen the emergence of innovative techniques that enable personalized explanations. These tools, such as **SHAP (Lundberg & Lee, 2017)** and **LIME (Ribeiro, Singh, & Guestrin, 2016; 2018)**, estimate the importance of each feature at the level of individual predictions, even for complex or black-box models.

On the other hand, **global interpretability** aims to give a comprehensive overview of how a model behaves across its entire input space. It seeks to provide a high-level understanding of the model's logic, structure, and decision patterns (Du et al., 2019). Achieving global interpretability requires access to a trained model, as well as an understanding of the algorithm and data used to build it (Lipton, 2016). This type of explanation is particularly useful when users need to trust the model's overall behavior rather than individual predictions.

An additional perspective within this area is known as **cohort-specific interpretability**, as described by Ahmad, Eckert, Teredesai, and McKelvey (2018). This approach involves analyzing subgroups within a population to understand how group-specific traits influence model outcomes. These subgroup-level explanations may be viewed as **global** if the cohort is treated as a subpopulation, or as **local** if they

represent grouped individual-level explanations (Molnar, 2020).

One notable technique in this space is **Model Understanding through Subspace Explanations (MUSE)**, which explains predictions based on feature-defined subspaces within the data (Lakkaraju, Kamar, Caruana, & Leskovec, 2019). MUSE highlights how particular features define subgroups, and how those subgroups behave in relation to the model's predictions, offering a hybrid of local and global insights.

TABLE 1Examples of approaches to interpretability of prediction regression models

	Global	Local
Model-specific	- Decision trees (depends on depth and number of terminal nodes; Hastie, Tibshirani, & Friedman, 2009; Stiglic, Kocbek, Pernek, & Kokol, 2012), - Linear and logistic regression models (Harrell Jr, 2015), - Generalized linear models (GLM) and generalized additive models (GAM; Hastie et al., 2009), - Naive Bayes classifier (Kononenko, 1993), - GNNExplainer (Ying et al., 2019)	- Set of rules (for specific individual; Visweswaran, Ferreira, Ribeiro, Oliveira, & Cooper, 2015), - Decision trees (by tree -decomposition; Visweswaran et al., 2015), - Visual analytics-based approaches (interactive visualization techniques for interpretation focusing on individual prediction), - k-Nearest neighbors (k-NN; depends on the number of important features, retrieving k-nearest neighbors for interpretation; Yuwono et al., 2015), - GNNExplainer (Ying et al., 2019)
Model-agnostic	- Different variants of model compression/knowledge distillation/global surrogate models (Elshawi, Al-Mallah, & Sakr, 2019), - Partial Dependence Plots (PDP; Elshawi, Al-Mallah, & Sakr, 2019), - Individual Conditional Expectation (ICE) plots (Elshawi, Al-Mallah, & Sakr, 2019), - Black Box Explanations through Transparent Approximations (BETA; Lakkaraju, Kamar, Caruana, & Leskovec, 2017) - Model understanding through subspace explanations (MUSE; Lakkaraju et al., 2019).	Local interpretable model-agnostic explanations (LIME; Ribeiro et al., 2016), Shapley additive explanations (SHAP; Lundberg & Lee, 2017), Anchors (Ribeiro et al., 2018), Attention map visualization, Model understanding through subspace explanations (MUSE; Lakkaraju et al., 2019).

In contrast, there are scenarios where understanding model predictions requires a **local interpretability** approach. One notable advantage of **model-specific local methods** is their capacity to explain individual decisions in a clear and intuitive manner—often through a concise set of rules, a segment of a decision tree, or by showcasing a few examples similar to the instance being evaluated (e.g., nearest neighbors). However, these methods typically trade off **predictive accuracy** for interpretability, performing less effectively than **model-agnostic techniques**, which often involve more complex algorithms. Moreover, many techniques in this category offer only limited

interpretative depth—commonly restricted to **feature importance-based explanations** for individual predictions..

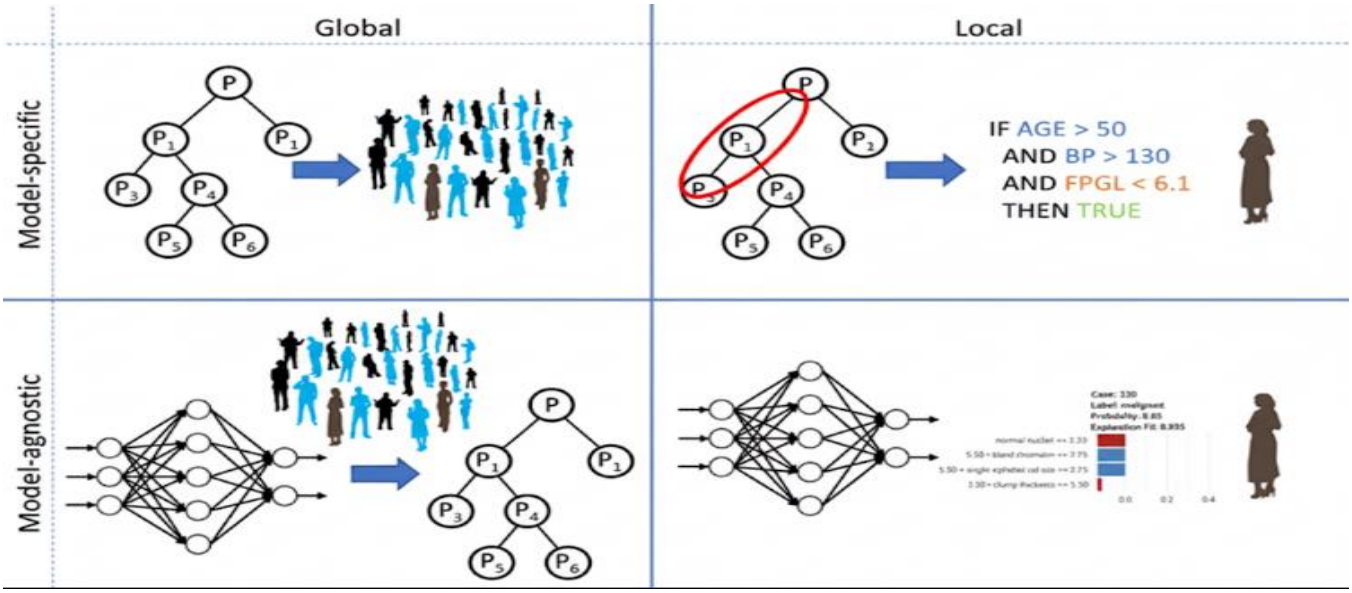
Practical Use of Interpretable Machine Learning in Healthcare

This section highlights real-world healthcare studies that have implemented one or more interpretability methods outlined earlier, showcasing their utility across various medical domains.

Model-specific techniques that emphasize **global interpretability** have been applied in the healthcare field for over twenty years. Their continued use is largely due to their transparency and ease of application in clinical practice. Algorithms like **linear regression** and **naive Bayes** remain prevalent in disciplines such as **urology** (Otunaiya & Muhammad, 2019; Zhang et al., 2019), **toxicology** (Zhang et al., 2018; Zhang et al., 2019), **endocrinology** (Alaoui, Aksasse, & Farhaoui, 2019), **neurology** (Zhang & Ma, 2019), **cardiology** (Doshi-Velez & Kim, 2018; Feeny et al., 2019; Salmam, 2019), and **psychiatry** (Guimarães et al., 2019; Obeid et al., 2019). Despite their interpretability, these models face limitations—particularly when dealing with **nonlinear relationships** or **heterogeneous data**, which can make their outputs less straightforward to interpret. On the other hand, **local interpretability methods** based on **model-specific techniques** such as **k-nearest neighbors (k-NN)** and **decision trees** have more recently been applied to explain predictions in areas like **occupational health conditions** (Di Noia et al., 2020) and **knee osteoarthritis** (Jamshidi et al., 2019). These methods have also supported the interpretation of ML-based predictions in **oncology**, including cancers such as **breast and prostate cancer** (Aro et al., 2019; Seker et al., 2000), as well as in determining **disease severity** for chronic illnesses like **diabetes** and **Alzheimer's disease** (Bucholc et al., 2019; Karun et al., 2019). Additionally, they have been employed in predicting **mortality risks** related to conditions such as **myocardial infarction** or **perinatal stroke** (Gao et al., 2020; Prabhakararao & Dandapat, 2019).

FIGURE 1: Visual representation of interpretability approaches for machine learning-based predictive modeling in healthcare.

Local and model-agnostic interpretability techniques are particularly useful when dealing with complex machine learning models like deep learning. For instance, the SHAP method was applied in a study to interpret predictions for preventing hypoxemia during surgical procedures. This application led to a **15% improvement** in anaesthesiologists' ability to anticipate hypoxemia events (Lundberg et al., 2018).As discussed earlier, some recent interpretability methods do not fit neatly into the four categories illustrated in Figure 1. One such method is **Model Understanding through Subspace Explanations (MUSE)**, introduced by Lakkaraju et al.



(2019), which uses expert-defined feature subspaces for interpretation. MUSE was implemented on a neural network trained for diagnosing depression. It generated if-then rule sets that offered a global explanation of the model's behavior, while also providing separate rule sets tailored to specific subspaces chosen by healthcare experts. For example, the model might focus on features such as **exercise habits** or **smoking**, which are considered actionable and modifiable by patients. A standard global interpretability technique typically treats all features equally and does not emphasize those with the potential for intervention. MUSE, therefore, although generally categorized under **global model-agnostic interpretability**, also integrates **personalized, user-driven interpretation**, making it a hybrid approach.Beyond technical considerations in building interpretable models, it is equally important to consider **ethical, legal, and regulatory** aspects. For example, the **General Data Protection Regulation (GDPR)** enforces a "right to explanation" (Wachter et al., 2017; Wallace & Castro, 2018). This has led to a shift toward more **user-centric** interpretability methods, enabling users to select the most suitable model for their needs. If the goal is to understand predictions at a broader, population level, then **global and model-specific techniques** might be the best fit. However, most real-world healthcare applications require a **personalized approach**, making **local and model-specific interpretability** more appropriate for providing detailed, patient-level insights.That said, a challenge remains: while local model-specific methods (like **personalized rule sets** or **k-nearest neighbors)** offer greater interpretability, they often fall short in predictive accuracy compared to advanced techniques such as **deep learning**, which generally do not support model-specific interpretability.

Discussion

This paper offers a timely and actionable review of interpretable methods for predictive models in healthcare, emphasizing what is practical from the perspective of those who will use them. Healthcare decisions are inherently high stakes, able to affect treatment outcomes or even survival. Our focus has been on structured data, where much of the attention is on which features (variables) matter most. But interpretability is also critical with unstructured data in fields like medical computer vision and natural language processing.

In medical imaging (computer vision), for example, methods seek to highlight the parts of an image that drive important health-related diagnoses, so clinicians can see which image regions are triggering model predictions. MDPI In NLP, interpretability often means marking or extracting segments of text that explain why a document was categorized in a particular way; this applies when analysing clinical notes or electronic health record (EHR) text.

Over recent years, there has been a shift away from globally interpretable, model-specific techniques toward more complex models that require **local, model-agnostic** methods. One reason is that massive datasets are now available in healthcare, making complex models more feasible and often necessary.

There remain multiple open challenges. Simplifying complex models for explanation can lead to loss of nuance or suboptimal performance. Auditing for fairness and bias is crucial. Methods like LIME or SHAP can be computationally expensive when applied at scale. Also, evaluating explanation methods themselves is still under development—two models may achieve similar accuracy but offer very different explanations. Finally, there is a growing interest in causal interpretability—understanding not just **what** features correlate with outcomes, but **why** those relationships exist.

Box 1: Visual Analytics and Interpretability

Visual analytics (VA) techniques are gaining traction in healthcare as tools that combine visualization and interactive feedback to enhance interpretability of machine learning models. The VA paradigm goes beyond static visualizations by allowing domain experts to interact with model outputs and steer the model using their feedback.

For instance, **RetainVis** is a visual analytics tool that works with recurrent neural networks (RNNs) on electronic medical records. It enables users to see how individual medical codes and visits contribute to risk predictions. It also integrates temporal information (when events happened) and allows what-if analyses—such as adding or removing codes or altering time

intervals between visits—to see how predictions change. arXiv+2PubMed+2

Many VA tools are model-agnostic. **RuleMatrix**, for example, lets domain experts see rule-based summaries of black-box model behavior to help inspect and understand model predictions. **Prospector** offers both population-level and individual instance explanations by exposing how input features jointly affect predictions. VA systems also make use of visualizations like heatmaps, attention maps, t-SNE projections, etc., to help make sense of complex models.

Conclusion

To build trust and ensure fairness and transparency in machine learning (ML) predictions, especially in healthcare, it is vital to understand various interpretability strategies. These can be broadly classified as either **model-specific or model-agnostic**, and as **global or local** interpretability methods. Although a wide range of interpretability techniques is now available, many core challenges remain unresolved. Addressing these will require continued research and the development of innovative solutions.

A major obstacle arises from the increasing complexity of predictive models. For instance, integrating medical knowledge through **knowledge graphs** introduces new interpretability challenges, particularly in how results are represented and understood. Combining different interpretability methods—like those discussed throughout this article—may offer more effective solutions.

Looking ahead, we anticipate broader adoption of hybrid approaches such as **MUSE** (Lakkaraju et al., 2019), which complement global interpretability with tailored insights for individual patients or specific patient subgroups. While the future direction of interpretability research is still evolving, it is clear that this topic is fundamental to the responsible development of prediction models in healthcare.

Another promising area is **Graph Neural Networks (GNNs)**, which leverage both node features and the structure of graphs through neural networks. Tools like **GNNExplainer** (Ying et al., 2019) offer potential for post hoc explanation of GNNs, yet our current understanding of their behavior and limitations is limited (Xu et al., 2018). Improving the **generalization** abilities of these models is an important next step.

Ultimately, future research should prioritize algorithmic advances that empower machine learning to support critical healthcare decisions—particularly those affecting disease progression, treatment outcomes, and personalized care strategies.

Acknowledgment

The author(s) gratefully acknowledge the financial support provided by the Slovenian Research Agency through the research grants **ARRS N2-0101** and **ARRS P2-0057**.

Conflict of Interest

The author(s) confirm that there are no competing interests or conflicts of any kind related to the content or publication of this work.

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: Survey on explainable artificial intelligence (XAI). In IEEE access (pp. 52138–52160). New York, NY: IEEE.

Ahmad, A. M., Eckert, C., Teredesai, A., & McKelvey, G. (2018). Interpretable machine learning in healthcare. In IEEE intelligent informatics bulletin (pp. 1–7). New York, NY: IEEE.

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. pp. 559–560.

Alaoui, S. S., Aksasse, B., & Farhaoui, Y. (2019). Data mining and machine learning approaches and Technologies for Diagnosing Diabetes in women. In International Conference on Big Data and Networks Technologies. Springer, Cham. pp. 59–72.

Aro, T. O., Akande, H. B., Jibrin, M. B., & Jauro, U. A. (2019). Homogenous ensembles on data mining techniques for breast cancer diagnosis. Daffodil International University Journal of Science and Technology, 14(1), 9–12.

Arras, L., Horn, F., Montavon, G., Müller, K. R., & Samek, W. (2017). "What is relevant in a text document?": An interpretable machine learning approach. PLoS One, 12(8), e0181142.

Bibal, A., & Frenay, B. (2016). Interpretability of machine learning models and representations: An Introduction. In 24th European symposium on artificial neural networks, computational intelligence and machine learning, Bruges. pp. 77–82.

Bratko, I. (1997). Machine learning: Between accuracy and interpretability. In G. Della Riccia, H.-J. Lenz, & R. Kruse (Eds.), Learning, networks and statistics (pp. 163–177). Vienna: Springer.

Breiman, L. (2001). Statistical modelling: The two cultures (with comments and a rejoinder by the author). Statistical Science, 16(3), 199–231.

Bucholc, M., Ding, X., Wang, H., Glass, D. H., Wang, H., Prasad, G., ... KongFatt, W. (2019). A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. Expert Systems with Applications, 130, 157–171.

Bucila, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. In KDD '06 Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY. pp. 535–541.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8 (832), 1–34.

Di Noia, A., Martino, A., Montanari, P., & Rizzi, A. (2020). Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction. Soft Computing, 24(6), 4393–4406.

Doshi-Velez, F., & Kim, B. (2018). Considerations for evaluation and generalization in interpretable machine learning. In H. Jair Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. A. J. van Gerven (Eds.), Explainable and interpretable models in computer vision and machine learning (pp. 3–17). Cham: Springer.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. Communications of the ACM, 63(1), 68–77.

Elshawi, R., Al-Mallah, M., & Sakr, S. (2019). On the interpretability of machine learning based model for predicting hypertension. BMC Medical Informatics and Decision Making, 19(1), 146.

Endert, A., Ribarsky, W., Turkay, C., Wong, B. W., Nabney, I., Blanco, I. D., & Rossi, F. (2017). The state of the art in integrating machine learning into visual analytics. Computer Graphics Forum, 36(8), 458–486.

Escalante, H. J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., & Güçlü, U. M.. (2018). Explainable and interpretable models in computer vision and machine learning. Cham: Springer International Publishing.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25, 24–29.

Feeny, A. K., Rickard, J., Patel, D., Toro, S., Trulock, K. M., Park, C. J., ... Gorodeski, E. Z. (2019). Machine learning prediction of response to cardiac resynchronization therapy: Improvement versus current guidelines. Circulation. Arrhythmia and Electrophysiology, 12(7), e007316.

Gao, Y., Long, Y., Guan, Y., Basu, A., Baggaley, J., & Plötz, T. (2020). Automated general movement assessment for perinatal stroke screening in infants. In F. Chen, R. I. García-Betances, L. Chen, M. F. Cabrera, & C. D. Nugent (Eds.), Smart assisted living (pp. 167–187). Cham: Springer.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. In Fifth International Conference on Data Science and Advanced Analytics (DSAA). New York, NY: IEEE. pp. 80–89.

Greene, T., Shmueli, G., Ray, S., & Fell, J. (2019). Adjusting to the GDPR: The impact on data scientists and behavioral researchers. Big Data, 7(3), 140–162.

Guimar˜aes, A. J., Araujo, V. J. S., Araujo, V. S., Batista, L. O., & de Campos Souza, P. V. (2019, May). A hybrid model based on fuzzy rules to act on the diagnosed of autism in adults. In IFIP International Conference on Artificial Intelligence Applications and Innovations. Cham: Springer. pp. 401–412.

Hall, P., & Gill, N. (2018). An Introduction to machine learning interpretability: An applied perspective on fairness, accountability, transparency, and explainable AI. Boston, MA: O'Reilly.

Hall, P., Gill, N., Kurka, M., & Phan, W. (2019). Machine learning interpretability with H_2O driverless AI. Mountain View, CA: H2O.ai, Inc.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), Advances in neural information processing systems (pp. 1024–1034). Cambridge, MA: MIT Press.

Harrell, F. E., Jr. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis. Cham: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Berlin: Springer.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. NIPS Deep Learning and Representation Learning Workshop.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainabilty of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9, e1312.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. (2018). Artificial intelligence in radiology. Nature Reviews Cancer, 18 (8), 500–510.

Jamshidi, A., Pelletier, J. P., & Martel-Pelletier, J. (2019). Machine-learning-based patient-specific prediction models for knee osteoarthritis. Nature Reviews Rheumatology, 15(1), 49–60.

Jia, X., Ren, L., & Cai, J. (2020). Clinical implementation of AI technologies will require interpretable AI models. Medical Physics, 47(1), 1–4.

Karun, S., Raj, A., & Attigeri, G. (2019). Comparative Analysis of Prediction Algorithms for Diabetes. In S. K. Bhatia, S. Tiwari, K. K. Mishra, & M. C. Trivedi (Eds.), Advances in computer communication and computational sciences (pp. 177–187). Singapore: Springer.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17, 195.

Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. Applied Artificial Intelligence an International Journal, 7(4), 317–337.

Kovalerchuk, B., Vityaev, E., & Ruiz, J. F. (2001). Consistent and complete data and "expert" mining in medicine. Studies in Fuzziness and Soft Computing, 60, 238–281.

Krause, J., Perer, A., & Bertini, E. (2016). Using visual analytics to interpret predictive machine learning models. arXiv preprint arXiv: 1606.05685.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2017). Interpretable & explorable approximations of black box models. arXiv preprint arXiv:1707.01154.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. In AIES '19 Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY: ACM. pp. 131–138.

Lei, T. (2017). Interpretable neural models for natural language processing (doctoral dissertation). Cambridge, MA: Massachusetts Institute of Technology.

Li, Y., Fujiwara, T., Choi, Y. K., Kim, K. K., & Ma, K. L. (2020). A visual analytics system for multi-model comparison on clinical data predictions. arXiv preprint arXiv:2002.10998.

Lipton, Z. C. (2016). The mythos of model interpretability. arXiv preprint arXiv:1606.03490.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus Explainability. The Hastings Center Report, 49 (1), 15–21.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., ... Lee, S. I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering, 2(10), 749–760.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579–2605.

Martens, D., Huysmans, J., Setiono, R., Vanthienen, J., & Baesens, B. (2008). Rule extraction from support vector machines: An overview of issues and application in credit scoring. In J. Diederich (Ed.), Rule extraction from support vector machines (pp. 33–63). Berlin, Heidelberg: Springer.

Mazurowski, M. A., Buda, M., Saha, A., & Bashir, M. R. (2019). Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. Journal of Magnetic Resonance Imaging, 49(4), 939–954

Michalopoulos, G., Chen, H., Yang, Y., Subendran, S., Quinn, R., Oliver, M., ... Wong, A. (2020). Why do I trust your model? Building and explaining. Predictive models for peritoneal dialysis eligibility. Journal of Computational Vision and Imaging Systems, 5(1), 1.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.

Ming, Y., Qu, H., & Bertini, E. (2018). Rulematrix: Visualizing and understanding classifiers with rules. IEEE Transactions on Visualization and Computer Graphics, 25(1), 342–352.

Molnar, C. (2020). Interpretable machine learning: A guide for making black box models explainable. Victoria, BC, Canada: Leanpub.com.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2018). Interpretable machine learning: Definitions, methods, and applications. Proceedings of the National Academy of Sciences, 116(44), 22071–22080.

Nebrini, S. (2019). Bias in the intervention in prediction measure in random forests: Illustrations and recommendations. Bioinformatics, 35 (13), 2343–2345.

Khan, T. (2023). AN AGRICULTURAL INTERNET OF THINGS (A-IOT) BASED INTELLIGENT SYSTEM FOR DISEASE PREDICTION USING TRANSFER LEARNING, A CASE STUDY. Lahore Garrison University Research Journal of Computer Science and Information Technology, 7(3).

Obeid, J. S., Weeda, E. R., Matuskowitz, A. J., Gagnon, K., Crawford, T., Carr, C. M., & Frey, L. J. (2019). Automated detection of altered mental status in emergency department clinical notes: A deep learning approach. BMC Medical Informatics and Decision Making, 19(1), 164.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

Otunaiya, K. A., & Muhammad, G. (2019). Performance of Datamining techniques in the prediction of chronic kidney disease. Computer Science and Information Technology, 7(2), 48–53.

Piltaver, R., Luštrek, M., Gams, M., & Martinc˘ic-Ipšic, S. (2016). What makes classification trees comprehensible? Expert Systems with Applications, 62, 333–346.

Abbas, T., Fatima, A., Shahzad, T., Alharbi, M., Khan, M. A., & Ahmed, A. (2024). Multidisciplinary cancer disease classification using adaptive FL in healthcare industry 5.0. Scientific Reports, 14(1), 18643.

Janjua, J. I., Nadeem, M., Khan, Z. A., & Khan, T. A. (2022, April). Computational Intelligence Driven Prognostics for Remaining Service Life of Power Equipment. In 2022 IEEE Technology and Engineering Management Conference (TEMSCON EUROPE) (pp. 1-6). IEEE.

Al-Tarawneh, A. M., AlOmoush, R. A., ul Islam, T., Janjua, J. I., Abbas, T., & Ihsan, A. (2024, December). Current Trends in Artificial Intelligence for Educational Advancements. In 2024 International Conference on Decision Aid Sciences and Applications (DASA) (pp. 1-6). IEEE.

Alqarafi, A., Batool, H., Abbas, T., Janjua, J. I., Ramay, S. A., & Ahmed, M. (2024, December). Estimating Uncertainty in Deep Learning Methods and Applications. In 2024 International Conference on Computer and Applications (ICCA) (pp. 1-6). IEEE.

Ramay, S. A., Kanwal, K., Javid, H. A., Abbas, T., Ansari, G. J., & Irfan, M. (2023). Enhancing Fruit Quality Detection with Deep Learning Models. Journal of Computing & Biomedical Informatics, 6(01), 28-40.

Prabhakararao, E., & Dandapat, S. (2019). A weighted SVM based approach for automatic detection of posterior myocardial infarction using VCG signals. In 2019 National Conference on Communications (NCC). New York, NY: IEEE. pp. 1–6.

Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. In N. Dey, A. Ashour, & S. Borra (Eds.), Classification in BioApps (pp. 323–350). Cham: Springer.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. In Proceedings of the 2016 ICML workshop on human interpretability in machine learning (WHI 2016). pp. 91–95.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In Thirty-Second AAAI Conference on Artificial Intelligence.

Salmam, I. (2019). Heart attack mortality prediction: An application of machine learning methods. Turkish Journal of Electrical Engineering and Computer Sciences, 27(6), 4378–4389.

Seker, H., Odetayo, M. O., Petrovic, D., Naguib, R., & Hamdy, F. (2000). A soft measurement technique for searching significant subsets of prostate cancer prognostic markers. In P. Sincak, J. Vascak, V. Kvasnicka, & R. Mesiar (Eds.), The state of the art in computational intelligence (pp. 325–328). Heidelberg: Physica.

Simon, G. E., Shortreed, S. M., & Coley, R. Y. (2019). Positive predictive values and potential success of suicide prediction models. JAMA Psychiatry, 76(8), 868–869.

Simpao, A. F., Ahumada, L. M., Gálvez, J. A., & Rehman, M. A. (2014). A review of analytics and clinical informatics in health care. Journal of Medical Systems, 38(4), 45.

Steyerberg, E. W. (2019). Clinical prediction models. Cham: Springer International Publishing.

Stiglic, G., Kocbek, S., Pernek, I., & Kokol, P. (2012). Comprehensive decision tree models in bioinformatics. PLoS One, 7(3), e33812.

Stiglic, G., Mertik, M., Podgorelec, V., & Kokol, P. (2006). Using visual interpretation of small ensembles in microarray analysis. In 19th IEEE symposium on computer-based medical systems (CBMS'06). New York, NY: IEEE. pp. 691–695.

Tomasello, M. (1999). The cultural origins of human cognition. Cambridge, MA: Harvard University Press.

Ustun, B., & Rudin, C. (2017). Optimized risk scores. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (KDD).

van der Maaten L. (2018). Dos and Don'ts of using t-SNE to Understand Vision Models, CVPR 2018 Tutorial on Interpretable Machine Learning for Computer Vision. Retrieved from http://deeplearning.csail.mit.edu/slide_cvpr2018/laurens_cvpr18tutorial.pdf.

van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In Proceedings of the National Conference on Artificial Intelligence, San Jose, CA, 25–29 July 2004; AAAI Press: Menlo Park, CA; MIT Press: Cambridge, MA, pp. 900–907.

Vellido, A. (2019). The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Computing and Applications, 1–15.

Visweswaran, S., Ferreira, A., Ribeiro, G. A., Oliveira, A. C., & Cooper, G. F. (2015). Personalized modeling for prediction with decision-path models. PLoS One, 10(6), e0131022.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law, 7(2), 76–99.

Wallace, N., & Castro, D. (2018, March 26). The impact of the EU's new data protection regulation on AI. Retrieved from http://www2.datainnovation.org/2018-impact-gdpr-ai.pdf.

Wang, F., & Preininger, A. (2019). AI in health: State of the art, challenges, and future directions. Yearbook of Medical Informatics, 28(1), 16–26.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 25(9), 1337–1340.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks? arXiv preprint arXiv:1810.00826.

Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNN explainer: A tool for post-hoc explanation of graph neural networks. arXiv preprint arXiv:1903.03894.

Yuwono, T., Setiawan, N. A., Nugroho, A., Persada, A. G., Prasojo, I., Dewi, S. K., & Rahmadi, R. (2015). Decision support system for heart disease diagnosing using K-NN algorithm. Proceeding of the Electrical Engineering Computer Science and Informatics, 2(1), 160–164.

Zhang, H., Ma, J. X., Liu, C. T., Ren, J. X., & Ding, L. (2018). Development and evaluation of in silico prediction model for drug-induced respiratory toxicity by using naïve Bayes classifier method. Food and Chemical Toxicology, 121, 593–603.

Zhang, H., Ren, J. X., Ma, J. X., & Ding, L. (2019). Development of an in silico prediction model for chemical-induced urinary tract toxicity by using naïve Bayes classifier. Molecular Diversity, 23, 381–392.

Zhang, Y., & Ma, Y. (2019). Application of supervised machine learning algorithms in the classification of sagittal gait patterns of cerebral palsy children with spastic diplegia. Computers in Biology and Medicine, 106, 33–39.

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595.