# Chapter 6
# Explainable AI and Its Applications in Healthcare

**Arjun Sarkar**

## 6.1 Introduction

Due to the lack of high-end graphics or tensor processing units, previously, deep neural networks could not be implemented as state-of-the-art Artificial Intelligence (AI) algorithms. Rather, linear models were preferred, and they were easy to understand and interpret. Things started changing with the advent of more advanced processing units, in the last decade, when the algorithms took on real-world problems. The models began getting bigger and better. While this highly improved the model performances, this also led to a problem of model interpretability. With access to large datasets, such as ImageNet (Krizhevsky et al. 2017), and these larger non-linear models with millions of parameters, AI soon started taking on human performance, on certain tasks. In 2015, a deep learning model called ResNet (He et al. 2016) surpassed human accuracy at the ImageNet challenge. Soon, AI was implemented in real-world tasks, and companies, industries, and research facilities started adopting AI into their workflows.

AI has now become a part of our day-to-day lives. AI algorithms not only help with day to day tasks such as finding outlines in today's phone cameras or recommending movies on Netflix, but also take on more challenging tasks such as beating human beings at strategy games (Lee et al. 2016; Garisto 2019) and surpassing human beings in complex visual recognition tasks (Bartolo et al. 2020; Thompson and Baker 2021). The rise of deep learning algorithms (Lecun et al. 2015) and computational power over the years has led to this extreme advancement in AI.

Healthcare costs are on the rise all around the globe. AI in medicine and healthcare can reduce this cost and, at the same time, improve healthcare (Higgins and Madai

A. Sarkar (✉)
Leibniz Institute for Natural Product Research and Infection Biology, Hans Knöll Institute, Jena, Germany
e-mail: arjun.sarkar786@gmail.com; arjun.sarkar@leibniz-hki.de

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_6

2020). Even though we read online how an AI beats doctors at predicting a particular disease every other day, the reality of acceptance of AI is still low in healthcare. While there are many reasons for this, one of the most prevalent ones is the explainability of the AI model. The ‘black-box’ nature of deep learning models is yet to be fully understood, and this causes a lack of trust and transparency. One error by an AI algorithm can be fatal for a patient in a hospital. Thus, the healthcare sector is cautious about implementing AI without completely understanding these algorithms. Most AI software implemented in hospitals today only helps diagnose and aid the doctor in making decisions. The accepted AI software goes through many regulations before being implemented in a hospital environment.

To build trust and reliability on these ‘black-box’ models, a new research field has emerged in recent years—eXplainable Artificial Intelligence (XAI). This field focuses on interpreting AI models and aims to provide an understandable way to explain AI predictions. At the rise of the deep learning era, most of the research was focused on improving model performance without caring much about explainability. But, that trend is now changing, with many researchers and companies looking to provide high AI accuracies and increased interpretability of the AI models.

A question may arise here: Why can an AI with high accuracy be trusted blindly? Initially, since the AI gives high accuracy, it may seem that the model can be implemented in a real-world situation. But many studies have shown that AI does not always learn the things that humans want it to learn (Lapuschkin et al. 2019). In the PASCAL VOC challenge (Everingham et al. 2010), it was often noticed that the AI was not precisely detecting the object of interest but making its classification based on context (Lapuschkin et al. 2016). For example, a classifier was often noticed to predict images of horses based on the watermark on the pictures and not on the actual horses. Similarly, the algorithm predicted an image as a train, not based on the train itself but the railway tracks (Lapuschkin et al. 2016). So, even though the model gave good accuracy, the correct predictions were often based on some artifacts. Usually, people can’t comb through thousands of images on these big data challenges and figure out artifacts. So, these errors mostly go unnoticed. But these overfitting errors occur more often than expected. While this model trained on the PASCAL VOC dataset may perform overwhelmingly well on the test dataset, as the test data also belongs from the same distribution of images as the training data, the same model may fail miserably when tested on real-world data. This is just one of the many examples which fosters the need for the explainability of AI algorithms.

Sometimes, explainability is not about the end results but some intermediate learning. Deep learning algorithms have the power to find interesting patterns from images or text, which may be unknown to a human expert. When Deepmind’s AlphaGo AI defeated the rank one human, Lee Sedol, at the game of Go, it played certain moves that other Go experts termed as ‘not-human’ (Thompson and Baker 2021). This just meant that a human being would not make that move, or that move was previously unknown to humans. Similarly, these algorithms can find patterns in medical images or correlate specific genes with certain diseases previously unknown to health experts. In the scientific and healthcare field, this can prove to be revolutionary. Often scientists and doctors focus more on certain patterns and features than the final prediction, as those intermediate patterns can lead to new scientific discoveries. Due to the ability of deep learning algorithms to find patterns, these models have had massive success in the field of medicine (Ching et al. 2018; Piccialli et al. 2021), drug discovery (Chen et al. 2018; Gawehn et al. 2016), protein studies (Wang et al. 2017; Xu 2019), neuroscience (Marblestone et al. 2016; Richards et al. 2019), and radiology (Miotto et al. 2017; Kermany et al. 2018; Kuenzi et al. 2020).

The first part of the chapter looks at explainability from different aspects—the multidisciplinary nature of explainable AI in technological, legal, medical, and ethical aspects. Secondly, several explainability algorithms developed over the years which had significant impact on healthcare are explained in the next section. Finally, applications of these algorithms in real world medical tasks are showcased including the use of XAI in the recent COVID-19 pandemic.

## 6.2 The Multidisciplinary Nature of Explainable AI in Healthcare

The explainability of AI in the healthcare domain is not always a technological issue. It can often be due to combined medical, legal, or ethical issues (Amann et al. 2020).

### *6.2.1 Technological Outlook*

The main issue of XAI is a technological problem: trying to explain an AI algorithm in a human-understandable form. The AI algorithm itself can achieve this explainability, or different models or methods can be used to describe a trained model (Rudin 2019). While the former can be achieved easily for linear models, the latter is necessary for the larger and more complex deep learning models.

Since the inception of XAI, various methods have been developed to try and explain these deep learning models. The explanation of linear models is always very accurate. But these models have severe performance issues compared to the more complex AI models (Esteva et al. 2019). So, there is a tradeoff between the complexity of a model and its explainability. Not only does model understanding help in understanding the final decision, but it also aids developers in tuning the parameters of the model and increasing performance. The problems of overfitting can be reduced or removed altogether. Researchers at Mount Sinai hospital trained a deep learning model to classify safe and high-risk patients based on X-ray images (Zech et al. 2018). The model produced high accuracies on the test set. But when the same model was tested on hospitals other than Mount Sinai, the model performance decreased. When XAI techniques were applied to the model, it was noticed that the model learned from the metadata of the X-ray machine at Mount Sinai hospital rather than on the actual X-ray images. The model was thus able to distinguish the pictures easily from that particular X-ray machine but failed on images of other X-ray machines, as the metadata no longer matched. XAI techniques help identify and correct these problems before the model is deployed in a real-world scenario. This makes the model more robust, reduces integration costs, and saves time.

While certain XAI techniques help developers improve model parameters, other techniques help healthcare professionals without in-depth knowledge in programming understand predictions. Pointing out the position of infection in medical images has immense benefits for doctors and helps provide a second opinion when they are in doubt. AI also can find rare diseases that are not often known to even seasoned experts (Schaefer et al. 2020). During the recent COVID-19 pandemic, while many algorithms were developed to classify whether patients were infected or not, very few were deployed on the field due to a lack of proper explainability (Fuhrman et al. 2022).

### *6.2.2 Legal Outlook*

The explainability of AI in healthcare is a legal need in nearly every country. In different sectors, the legal requirement for XAI is dissimilar. XAI is not a must in logistics, and a few errors are admissible. But in public administration or banking, XAI can play a vital role. A person whose loan has been rejected due to an AI model has the right to know the reason behind the rejection. In no other sector is XAI as mandatory as in the healthcare field (Schönberger 2019). This does not come as a surprise, as in healthcare, even one error has the potential of harming human life.

AI in healthcare is used for many applications, such as disease classification and diagnosis (Qiu et al. 2020), anomaly detection, patient positioning, image segmentation (Aslam et al. 2015), image super resolution (Chaudhari et al. 2018), and image registration (Ma et al. 2017; Wu et al. 2013). AI is meant to improve clinical applications and aid doctors, improve the standard of medical development and save patient lives. But to train a robust model, often sensitive patient data is required. These privacy issues must meet all legal requirements, from image acquisition to final prediction. Similarly, in recent years, anti-discrimination and explainability of Ai models have gained momentum (Deeks 2019).

Hospitals don't use the AI algorithm as a computer program, but the algorithm is wrapped in the form of a software with a user-friendly graphical user interface (GUI). It is a requirement by most regulatory bodies in the USA or the European Union to provide a level of transparency of the AI's output (Smith et al. 2020). Though these regulations are rather vague now, with no solid rules for explainability, these rules will supposedly get stricter as more emphasis is made on XAI.

One more budding question is about the awareness of these AI predictions and the disclosure to patients. That is, how much of the decision would be made by the AI and how much by the doctor, and finally, how the final prediction would be disclosed to the patient (Cohen 2020). One fear is that the legal system is not fast enough to keep with the rising pace of AI development. In healthcare, AI-based decisions need strict laws such that they do not hamper innovation but also protect patients' rights and privacy. When these laws are clearly defined and AI researchers can overcome the problems with XAI, AI will be fast adopted in all healthcare sectors.

### *6.2.3 Medical Outlook*

The medical outlook aims to bring semblance between the need for laboratory-based testing or replacing it entirely with AI-based algorithms. Laboratory testing and medical imaging are the methods that have been used since time immemorial for the proper diagnosis of a disease. These are methods that are understandable by medical experts. In laboratory testing and imaging, doctors access the results and the images and find meaningful patterns that point to certain complications or diseases. On the other hand, when trained on the images and the corresponding infection types, a deep learning model may predict the condition correctly but does not indicate the patterns it uses to come to that prediction. XAI can be helpful here in showing these patterns and can be much faster than even the most trained experts.

Even though an AI algorithm can provide good overall accuracy and low error rates (Weng et al. 2017), the algorithm cannot be perfect because of data inconsistency due to noise and imaging errors. The trained experts need to look at the false positives and negatives, and they cannot always be heavily reliant on AI. AI bias is another such complication that cannot be removed entirely. For example, suppose the training data is sampled from a large population of people from Europe in skin cancer prediction. In that case, the same algorithm will fail if deployed in Africa due to the differences in skin tone and color (Wen et al. 2022).

XAI can be crucial in deciding the amount of disagreement between a medical expert and an AI. The results of XAI in the medical field are often visual representations or textual explanations. These explanations can be beneficial to the medical experts in making a final decision on the diagnosis. Without XAI, the clinician has to choose blindly whether to trust the AI or not, but with XAI, the person can understand why the AI makes that particular decision. If an algorithm keeps performing poorly, the results can be reported to the developers, and the developers can understand the reason for the poor performance using explainability. In case the algorithm works well, the clinicians can trust it better when they understand the reason for its good performance (Cutillo et al. 2020).

### *6.2.4 Ethical Outlook*

As more and more healthcare institutes adopt AI into their framework, certain ethical aspects need to be examined. One of these issues is protecting the autonomy of the patient and informing the patient about the use of the deep learning models for their diagnosis. If a patient is not informed whether a doctor or an AI algorithm predicted a particular disease, this can hamper the patient's trust towards the doctor. A more critical situation would be if the prediction is a false positive or false negative, and the patient is mistreated. The patient can challenge the institution, and the healthcare facility will not be able to provide a concrete reason for such a prediction. This is one more reason for introducing a proper XAI before using the AI algorithm blindly. A solution to such a scenario can be first asking the patient for their permission to use AI for the diagnosis and later explaining the results of the AI to them.

One more ethical conundrum can arise with the rise of AI in healthcare. If AI systems start taking over more healthcare positions in the future, it can limit clinicians' decision-making rather than enhance them. This situation should be avoided, as the best outcome for a patient is to be not entirely dependent on an AI, but a decision of AI carefully analyzed by a doctor.

### *6.2.5 Patient Outlook*

The patient outlook is a perspective that focuses on patients and considers them as an active part of the healthcare decision process (Baker 2001). This refers to the treatment process in which each patient is provided with a treatment best suited for that individual. The idea is to provide patient-centered medicine. But deep learning models predicting risk may not be able to provide such treatment. As doctors do not understand the inner workings of the models as well, neither can they inform the patient about the reason for the predicted risk. XAI can prove to be helpful and continue maintaining the patient-centered approach.

Similarly, wearable devices and smartwatches can now predict certain risk factors in patients (Bhattacharya and Lane 2016; Mauldin et al. 2018). Previously these devices provided similar treatment and health plans to all users. But recently, most companies have been trying to give each user a different health plan based on their activity, heart rate, and sleeping patterns (Coutts et al. 2020; Nweke et al. 2018). Risk assessment explained in the form of text or visual data builds trust and increases transparency and continues building towards a patient-centered innovation in healthcare.

## 6.3 Different XAI Techniques Used in Healthcare

There are various explainability methods for AI in medicine, and there are multiple ways to classify these methods. Some such taxonomies are explained below.

### 6.3.1 Methods to Explain Deep Learning Models

Since deep learning models are all black-box models and cannot be easily explained, most modern research is focused on trying to explain these models. Saliency maps is one such technique very commonly used to interpret convolutional neural networks (Itti et al. 1998). These saliency maps are pixels of the image that the convolutional neural network considers essential to the final prediction. Saliency is represented on the image as a visual heatmap or topography.

There are multiple gradient-based techniques used for the explainability of deep learning models (Simonyan et al. 2014). The gradient-based approach shows how much a change in the input would affect the output. Saliency maps are also based on this gradient technique. The Krizhevsky network (Krizhevsky et al. 2017) beat the previous methods and was considered one of the best gradient-based explainability methods. Another algorithm that improves gradient explainability is the DeepLIFT algorithm (Shrikumar et al. 2017). The DeepLIFT algorithm enhances the previous methods by multiplying the input signal to the gradient. The model's superiority was evident when tested on genomic data and natural images. The algorithm assigns a weighted score to the activation of all neurons in the model and shows some crucial connections or features that the previous models failed to identify. DeconvNets or Deconvolution (Zeiler and Fergus 2014) is another method to understand convolutional neural networks (CNN). Unlike regular convolutional layers that extract features from image pixels, deconvolution does the opposite—mapping features to pixel values. Deconvolution is generally used to understand what a convolutional neural network learns in every convolutional layer.

Class Activation Maps (CAM) (Zhou et al. 2016) and its more advanced counterparts Grad-CAM (Selvaraju et al. 2020) and Grad-CAM++ (Chattopadhay et al. 2018) are some of the most famous interpretability methods used to explain the results of convolutional neural networks. CAM helps to identify important locations of the image trained by a model to predict the class of the image. Activations from the final layer of the convolutional model are concatenated to create a feature vector. The weighted sum of this vector is fed into a SoftMax layer to calculate the final result. The result is displayed as a heatmap. While CAM gave good results, it could not be applied to any convolutional model. To overcome this problem, Grad-CAM was developed. Grad-CAM (Selvaraju et al. 2020) can generate the localization maps for any convolutional neural network, regardless of its shape or structure. Grad-CAM++ (Chattopadhay et al. 2018) further improves Grad-CAM by better visualization of the output maps and better object localization for multi-label classification.

The RISE algorithm (Petsiuk et al. 2019) slightly differs from the CAM algorithms. This algorithm considers each pixel of an image to generate a saliency map. Random masks are multiplied elementwise with the images and fed into the network. The model generates a probability score and a saliency map of the input image, which is obtained by combining the masks.

While many algorithms are trying to explain the results of the convolutional neural network models, some studies suggest that none of these techniques are correct in interpreting the networks (Kindermans et al. 2018). The authors tested some of these explainability methods on simple linear models, but the methods could not correctly interpret the linear models. Hence, the authors argue that if these techniques cannot even explain simple linear models, is their explanation of large complex non-linear models, correct? They further proposed two additional models, PatternNet and PatternAttribution, which work well on linear models and more complex models.

In Natural Language Processing (NLP), a different method is used for explainability (Lei et al. 2016). Small pieces of the input text are added to the model as input, and the model aims to generate the entire text from these small text fragments. Finally, the generated text provides some context and justification for the generated text in terms of the input text.

LIME (Ribeiro et al. 2016) or Local interpretable model-agnostic explanations is an XAI method that can interpret any black-box model. It is also one of the most famous and commonly used interpretability methods for tabular data, text, and images. LIME can interpret individual predictions of a model. It tweaks the feature values of a single data sample and creates an impact of the tweak on the output. Even though LIME can be a simple and powerful interpretable model, it has certain drawbacks. Some studies have shown that choosing poor parameters can cause the model to give different results and miss many essential features completely. This can be a severe problem when the model is deployed in the field. The DLIME algorithm was proposed to overcome this problem. Random sampling used in LIME is replaced in DLIME by choosing clusters of similar data and selecting the most relevant cluster by running k-nearest neighbors (KNN). The authors of DLIME also proved the superiority of DLIME over LIME by testing it on three separate medical datasets.

Shapely Additive explanations (SHAP) (Lundberg and Lee 2017) is another often used interpretability technique. SHAP is a model inspired by game theory. It computes the importance of each feature for all predictions. The SHAP values are a combination of three important properties, namely, accuracy, missingness, and consistency. The authors demonstrate how SHAP is more intuitive and more human interpretable than other XAI methods. Various other model agnostic models such as Anchors (Ribeiro et al. 2018), DeepSHAP (Chen et al. 2021), Protodash (Kim et al. 2016), Permutation Importance (PIMP) (Altmann et al. 2010), and Contrastive Explanation Methods (CEM) (Dhurandhar et al. 2018) are often used for explainability as well.

In deep learning, attention is a trendy concept (Vaswani et al. 2017). The concept of attention was inspired by how humans pay attention to different parts of images or other data sources to analyze them. The MDNet network was created (Xia et al. 2020) to directly map medical imaging and corresponding diagnostic reports. It contained an image model as well as a language model. Attention mechanisms were used to visualize the detection process. This attention mechanism allowed the language model to discover the predominant and distinguishing features used to map the images and diagnostic reports. This was the first study to use the attention mechanism to gain insight from the medical image data. SAUNet, an interpretable U-Net version (Ronneberger et al. 2015), was created (Sun et al. 2020). It also added a secondary shape stream to capture important shapes-based information in addition to the regular texture features. An attention module was used in the U-Net decoder. SmoothGrad (Hooker et al. 2019) was used to create spatially and shape attention mappings to visualize the high activation area of the images.

These are some commonly used XAI methods for deep learning models in healthcare. All these models have some significance, but there is no one idea to explain all kinds of text and image data or on all sorts of models. SHAP and its advancements are comprehensive and understandable XAI methods of all the methods. Grad-CAM is commonly used for convolutional model interpretation, even for industrial AI software deployed in hospitals.

### *6.3.2 Explainability by Using White-Box Models*

White-box models are transparent models and are easily understandable or interpretable. This category contains mainly linear models, decision trees, and some complex models that are easy to interpret. Some of these complex models used in medical imaging and healthcare are listed here.

Microsoft came up with an interpretable model for predicting pneumonia risk, which also had great accuracy (Caruana et al. 2015). The authors discussed that while interpretable linear models and decision trees could not give good results, neural nets gave much better results, but at the cost of explainability. High-performance generalized additive models with pairwise interactions (GA2Ms) were proposed and tested on two real medical data case studies. The authors also mentioned that the model could be scaled to work on thousands of patient data without losing accuracy and still being highly interpretable.

Another technique that utilizes Boolean rules to create predictive models was proposed—Boolean Rule Column Generation (Dash et al. 2018). This technique uses easy-to-understand Boolean rules with some clauses and conditions. Humans easily understand these clauses and conditions. GLM or Generalized Linear Rule Models (Wei et al. 2019) use an ensemble of rule-based features. GLMs are easy to interpret and complex simultaneously, as the rules can capture non-linear dependencies. TED or Teaching Explanations for Decisions (Hind et al. 2019) is a framework that tries to produce explanations like a human expert rather than explaining the inner workings of an AI model.

Not much research has been done in the complex white-box model development domain. No white-box model can produce the same high accuracy as deep learning models. The white-box models are also very domain-specific, unlike various computer vision and natural language processing neural networks used on various real-world tasks.

### *6.3.3 Explainability Methods to Increase Fairness in Machine Learning Models*

AI models are not just theoretical analysis techniques anymore, but with every passing day, more and more models are adopted in real-world applications. Any discrimination or inequality in these models can potentially impact human lives. The fairness of these AI models is another part of XAI that tackles ethical and social aspects. Usually, bias in the models is checked by implementing the model in a different setting, such as a different demographic, and evaluated. Many techniques developed in recent years focus on tackling the bias and discrimination in these models.

The method of disparate impact testing (Feldman et al. 2015) is a model-evaluation tool that can assess the fairness and accuracy of a model but does not provide any details or insight into the causes of bias. It uses simple experiments to highlight differences between model predictions and errors for different demographic groups. It can also detect biases in terms of ethnicity, gender, marital status, or demographics. Another data preprocessing technique was suggested to remove bias from machine-learning models (Calmon et al. 2017). The authors developed a convex optimization to learn a data representation that meets the three stated goals: controlling discrimination, limiting distortion in individual instances, and preserving utility. Adversarial debiasing (Zhang et al. 2018) is an approach to tackling biases regarding demographic segments in machine-learning systems. It involves selecting a feature about the element of interest and then simultaneously training both the primary and adversarial models. The main model is trained to predict the label. The adversarial model, based on the primary model's prediction for each instance, attempts to predict the segment. The goal is to maximize the main machine learning system's accuracy in correctly predicting the label while minimizing the adversarial ability. Adversarial biasing can be used for both classification and regression tasks.

Many methods to make classifiers aware of discriminatory biases need data modifications or algorithm tweaks (Kamiran et al. 2012). They are also not flexible regarding multiple sensitive features handling and control over performance versus discrimination tradeoff. Two new methods, Reject Option-based Classification and Discrimination-Aware Ensemble were developed to solve these problems.

Counterfactual fairness (Kusner et al. 2017) captures the intuition that a decision that affects an individual is fair if it affects the same person in both the real and counterfactual worlds. The individual would then be part of a different demographic. It was also argued that causality in fairness must be addressed. Consequently, a framework was developed to model fairness using tools of causal inference. The authors state that any measure of causality in fairness should not be based on counterfactuals. It is also essential to ensure that counterfactual causal guarantees can be used. Based on the concept of counterfactual fairness, the proposed framework allows users to create models that can take sensitive attributes that could reflect social biases towards people and compensate accordingly. A recent study (Kearns et al. 2018) found that most machine-learning fairness notions and definitions only focus on predefined social segments. It was also pointed out that while such simple constraints can force classifiers at the segment level to attain fairness, they could lead to discrimination against sub-segments that contain specific combinations of sensitive feature values. The authors suggested that fairness be defined across an infinite or exponential number of sub-segments. These were then determined using the space of sensitive feature values. An algorithm was developed to produce the fairest distribution of sub-segments over classifiers.

One study (Elisa Celis et al. 2019) pointed out that while recent research has attempted to attain fairness regarding a particular metric, specific metrics have been overlooked. Furthermore, some proposed algorithms lack solid theoretical support. The authors developed a meta-classifier that could handle multiple fairness constraints concerning multiple non-disjoint sensitive elements. Another work pointed out that many existing notions about fairness regarding treatment and impact are too restrictive and strict. This can lead to poor model performance. The authors suggested notions of fairness that were based on the collective preferences of different demographic groups to address this issue. Their concept of fairness, more specifically, tries to define which outcome or treatment the various demographic groups would prefer if given a choice.

Fairness is still a new area of machine learning interpretability. However, the incredible progress made over the past few years has been remarkable. Many methods can ensure fair resource allocation and protect the most vulnerable demographics. Several techniques can be used to manipulate data before training models, make algorithmic changes during training, and adjust post-hoc. However, these methods tend to focus too heavily on group fairness. They often overlook individual-level factors at both the local and global levels, leading to the mistreatment of individuals. A small portion of scientific literature deals with fairness in images or text. This gap is still a significant one that needs to be explored in the future.

### *6.3.4 Explainability Methods to Analyze Sensitivity of a Model*

Interpretability methods are used to evaluate and challenge machine learning models to ensure they are reliable and trustworthy. These methods use some form of sensitivity analysis. Models are evaluated for their stability and their ability to predict the impact of subtle but intentional changes in inputs. Sensitivity analysis may interpret changes in output across a range of examples or just one.

The sensitivity index is a traditional method of sensitivity analysis that represents each input variable using a numerical value. First-order indices measure the contribution of one input variable to the output variation. Second, third, and higher-order indexes measure the interaction contribution between two, three, or more input variables to that output variance. The total-effect indices combine the contributions of higher-order and first-order interactions with the output variance.

Sobol (2001) proposed an output variance sensitivity analysis based on ANOVA decomposition. He suggested using Monte-Carlo methods to approximate sensitivity indices higher and first order. Fourier Amplitude Sensitivity Test (FAST) (Cukier et al. 1973) is a method to improve the approximation of Sobol's indexes. The Fourier transformation converts a multi-dimensional integral to a one-dimensional integrated. These algorithms were further enhanced to an RBD-FAST (random balance designs-FAST) algorithm (Plischke 2010), which improved computational efficiency. Morris's method (Morris 1991) of global sensitivity analysis, also known as the one-step-at-a-time (OAT) method, is another option. Although the Morris method is complete, it can be very computationally expensive. Fractional factorial designs (Saltelli et al. 2008) needed to be developed and used in practice to perform sensitivity analysis more efficiently.

## 6.4 Application of XAI in Healthcare

There are two main types of explanations for deep neural networks in medical images: those that use standard attribution-based approaches and those that use novel, often domain-specific or architecture-specific methods. Many attribution methods can be used to assign an attribution value, contribution, or relevance to each network input feature. An attribution method determines the importance of an input element to the target neuron, which is often the output neuron for a classification problem. Heatmaps show the arrangement of all input features according to the shape of the input samples. Non-attribution is a methodology that is validated on an issue rather than using separate analyses using pre-existing methods. These included concept vectors, attention maps, return of a similar image, and text justifications.

Some applications of XAI in healthcare are explained in this section. While each healthcare domain has hundreds of studies where XAI has been used, only a few examples from each domain are listed.

### *6.4.1 Medical Diagnostics*

One study (Kavya et al. 2021) proposed a computer-aided framework for allergy diagnosis. They evaluated several ML algorithms and then chose the most effective one using k-fold cross-validation. They developed a rule-based approach to the XAI method by creating a random forest. If-Then rules and explanations representing each path within a tree are extracted using medical data. The computer-aided framework was also deployed on the mobile app by the authors, which can be used to assist junior clinicians in verifying the diagnostic predictions. Another study (Dindorf et al. 2021) suggests an explanation-independent classifier for spinal positions. SVM and radiofrequency were used as ML classifiers. Then, they applied LIME to predict the classification. The authors of another study (El-Sappagh et al. 2021) suggested an

RF model to diagnose and detect Alzheimer's progression. The authors also used SHAP to identify the essential features of the classifier. Next, they used a fuzzy rule-based method. SHAP could provide a local explanation for specific patient diagnosis/progression prediction explanations about feature impacts. The fuzzy rule-based system could also generate natural language forms that can aid patients and doctors in understanding the AI model. One paper suggested an XAI framework to assist doctors in diagnosing hepatitis patients (Peng et al. 2021). The authors compared intrinsic XAI methods such as logistic regression, decision trees, and kNN to the more complex models SVM, XGBoost, and RF. The authors also used the post-hoc methods SHAP and LIME and partial dependence plots (PDP). For chronic wound classification, a CNN model was proposed (Sarp et al. 2021). For XAI, the authors used LIME, which aided clinicians in better diagnosis.

## *6.4.2 Medical Imaging*

Due to their simplicity, attribution-based methods were used in most medical imaging literature. Researchers can efficiently train a neural network architecture that is suitable without making it difficult to explain. They also have access to an attribution model. A pre-existing deep learning model or a custom model can obtain the best results on a given task. The existing model implementation is more straightforward and can leverage transfer learning techniques. In comparison, custom models can concentrate on specific data and avoid overfitting with fewer parameters. Both are useful for medical imaging datasets.

Analyzing the post-model data using attributions can show if the model is learning the right features or if it's learning the wrong features. This allows researchers to adjust the hyperparameters and architecture of the model to get better results with test data and potentially in a real-world setting.

### 6.4.2.1 Brain Imaging

Different methods were analyzed in a study to compare their robustness in CNN's Alzheimer's classification using brain MRI. The methods that were compared were LRP (Bach et al. 2015) and Guided backpropagation (GBP). The L2 norm was calculated between the average attribution maps for multiple runs to check the repeatability of heatmaps of identically trained models. Because occlusion covers more area, it was an order of magnitude lower than the baseline occlusion. LRP performed better than all other methods, indicating a fully attribution-based method. LRP also had the highest similarity in the sum, density, and gain (sum/density), for the top 10 regions across all attributions. Another study (Pereira et al. 2018) used GradCAM and GBP to examine the clinical coherence between the features learned from a CNN for automatic grading brain tumors using MRI. Both methods activated the tumor and surrounding ventricles, which could indicate malignancy. They were both correctly graded in cases. This focus on non-tumor areas and spurious patterns in GBP maps can lead to errors that indicate unreliability.

#### 6.4.2.2 Breast Imaging

SmoothGrad and IG were used to visualize features in a CNN for classifying estrogen receptor status using breast MRI (Papanastasopoulos et al. 2020). The model learned relevant features from both dynamic and spatial domains, with each contribution. Visualizations showed that the model had learned some irrelevant features from pre-processing artifacts. These observations led us to make changes in our pre-processing and training methods. A previous study to classify breast mass from mammograms (Hassan et al. 2020) used two different CNNs, AlexNet (Szegedy et al. 2015) or GoogleNet (Krizhevsky et al. 2017)—and used saliency maps for visualizing image features. Both CNNs were able to detect the contours of the mass, which is the essential clinical criteria. They also showed sensitivity to context. In another study (Amoroso et al. 2021), the authors also presented an XAI framework to help breast cancer patients. The framework was used to identify a patient's most important clinical feature.

#### 6.4.2.3 Skin Imaging

GradCAM and KernelSHAP were used to compare the features of a set of 30 CNN models trained for melanoma detection (Young et al. 2019). GradCAM and Kernel SHAP were used to compare the features of a suite of 30 CNN models trained for melanoma detection. It was found that even high-accuracy models would sometimes focus on features that were not relevant to the diagnosis. The attribution maps of both methods showed differences in the models' explanations. This demonstrated that different neural network architectures learn various features. A further study (Molle et al. 2018) showed how CNN features were used to classify skin lesions. By scaling the feature maps of activations to the input size, the features for the two last layers were visualized. The layers looked for indicators such as lesion borders, color irregularities, and risk factors such as lighter skin or pink textures. However, some spurious features such as hair and artifacts had no significance.

#### 6.4.2.4 X-ray Imaging

Some studies have used attribution-based diagnostic methods in addition to the more popular imaging modalities. These include both image inputs and non-image inputs. One study used CNNs to perform uncertainty and interpretability analyses on colorectal polyps (Wickstrøm et al. 2020). This is a precursor to rectal cancers. CNN used GBP to create heatmaps. They were found to use the shape and edge information to make predictions. The uncertainty analysis also revealed higher levels of uncertainty in samples that were misclassified. The authors (Lundberg et al. 2018) presented a model that uses SHAP attributions for hypoxemia. This study was done to analyze preoperative and in-surgery factors. The resulting attributions were consistent with known factors such as BMI, physical state (ASA), tidal volumes, inspired oxygen, and others.

Attribution-based methods were the first method of visualizing neural networks. They have evolved from simple gradient-based class activation maps to more advanced techniques such as Deep SHAP. These visualizations show that models are learning relevant features in most cases. Any spurious features were flagged and corrected by the readers. The identification of relevant features can be improved by smaller models and custom variants to the attribution methods.

#### 6.4.2.5 CT Imaging

DeepDreams inspired attribution method (Mordvintsev et al. 2015) was presented in (Couteaux et al. 2019) to explain the segmentation and classification of tumors from liver CT images. Based on the DeepDreams concepts, this innovative method can be applied to black-box neural networks. The algorithm performed a sensitivity assessment of the features and maximized the activation of target neurons by performing gradient ascent. Comparing networks trained on synthetic and real tumors showed that the former was more sensitive than the latter to clinically relevant features. At the same time, the latter was also more focused on other features. The network was sensitive to both intensity and sphericity with domain knowledge.

#### 6.4.2.6 Retina Imaging

As a diabetic retinopathy (DR) tool, grading by ophthalmologists, a system that produced IG heatmaps and model predictions was investigated (Sayres et al. 2019). The assistance provided by the system was shown to improve the accuracy of the grading over that of an expert without any help or the model predictions. Although the initial grading process was slower, users soon found that it improved their grading experience. This is especially true when heatmaps and predictions are used. Patients without DR saw a decrease in accuracy when model assistance was used. Expressive gradients (EG) were proposed as an extension to IG for weakly supervised segmentation (Yang et al. 2019). Compact CNNs performed better than larger ones, and EG highlighted regions of interest more effectively than traditional IG or GBP methods. EG extends IG by enriching input-level attribution maps with high-level attribution maps. A comparative analysis of various explainability models, including DeepLIFT, DeepSHAP, IG, etc., was performed for a model for detection of choroidal neovascularization (CNV), and diabetic macular edema (DME), and drusens from optical coherence tomography (OCT) scans (Singh et al. 2020).

### *6.4.3 Surgery*

In one study (Kletz et al. 2019), the authors presented a CNN-based medical app to learn the representations of instruments in laparoscopy. They validated their model using different datasets. To help explain how the model classified an instrument, they also provided activation maps from different CNN layers. XAI-CBIR (Chittajallu et al. 2019) was proposed to explain surgical training. XAI-CBIR provides an example post hoc explanation of XAI methods. It extracts representative examples to offer explanations. It uses a self-supervised deep learning model to extract semantic descriptions from MIS video frames. It also used a saliency map to explain visually why it believes the image retrieved is similar to the query. Minimally invasive surgery (MIS) videos can be retrieved using the XAI CBIR system.

### *6.4.4 Detection of COVID-19*

Understanding the COVID-19 data associated with COVID-19 is necessary to fully understand the clinical applications of explainable AI for COVID-19 assessment (Fuhrman et al. 2022). While reverse transcription-polymerase chain reaction (RT-PCR) tests are the most common tool for COVID-19 detection, radiography, and CT scans can supplement RT-PCR testing to improve detection accuracy and throughput. For COVID-19 diagnosis, only X-ray or CT finding may not be sufficient. Therefore, differential diagnosis is difficult because of the subtle differences in COVID-19 and non-COVID-19 pneumonia (Cleverley et al. 2020). For improved differential diagnosis and detection accuracy, explainable AI can be helpful (Dong et al. 2021; Salehi et al. 2020). A standard language for describing COVID-19 can also be used. These datasets are publicly available and can meet data requirements. This includes large-scale projects such as the NIH-funded Medical Imaging and Data Resource Center.

This case study is mainly focused on radiography and chest CT. However, other modalities such as PET/CT, lung ultrasound, and MRI may also play a part in COVID-19 patient care. The development of AI systems to assess COVID-19 is similar to other disease evaluations in many ways. The most common use of explainable AI in COVID-19 assessment is to ensure the model accurately focuses on regions of concern in the input image that indicate disease presence. This is usually done through heatmap visualization. Some studies have had mixed success (Mei et al. 2020; Xiong et al. 2020; Wehbe et al. 2021). One study (Wehbe et al. 2021) shows heatmaps for both negative and positive COVID-19 cases. They note that the negative examples have a low influence on the lungs. Another study (Xiong et al. 2020) shows heatmaps that accurately highlight COVID-19 in lung segmentation and identify regions without significant COVID-19 content to aid the classification decision. This finding is limited in understanding the model's performance and should be investigated further before clinical implementation. COVID-19 can be easily confused with other diseases such as viral pneumonia. One study differentiates between these, and the results of this study are precise (Jin et al. 2020). The authors divided CT scans into four types of pneumonia and COVID-19. They also identified phenotypic mistakes that were common for humans and AI readers. Grad-CAM and Guided GradCAM were used to visualize the most critical image regions. The authors also provided segmentation for diseased areas. Like previous works, Grad-CAM indicates that the model identifies high-value regions within and outside of the lungs. However, Guided GradCAM does not capture all of the diseased lung tissue. They also use t-SNE to visualize feature embeddings from the various disease classes and identify image features that may be problematic in the classification decision. Another study (Zhang et al. 2020) presents another unique use of explainable AI in the COVID-19 assessment. In this case, the authors use clinical metadata and quantitative lesion features to create classifiers that can predict patient prognosis. They use Shapley numbers to assess how each feature impacts the risk classifier. This includes whether it increases or decreases a prediction output. They also evaluate the effectiveness of different drug administrations and the patient's response to treatment. This type of analysis is beneficial for understanding images indicative of high risk. It is helpful when combined with clinical metadata.

A method called GSInquire was used in a recent study to detect COVID-19 using chest X-ray images (Wang et al. 2020). It produced heatmaps that were used to verify the features of the COVID-net model. GSInquire was created to be an attribution method that performed better than other methods such as SHAP or Expected gradients. It uses the new metrics impact score and coverage. The impact score was the percentage of features that strongly impacted the model decision or confidence. Impact coverage was determined in relation to the inclusion of factors that could be adversely affected. While these studies use python programming to create the deep learning models, the Cognex VisionPro Deep Learning Software classified Covid-19 X-ray images using their deep learning-based graphic user interface (GUI) (Sarkar et al. 2021). The software has built-in Grad-CAM for interpretability, highlighting the regions of interest. A trained medical expert can then look at the Grad-CAM and judge the efficacy of the software.

## 6.5 Conclusion

Despite its rapid growth, explainable AI is still not a mature field. It often suffers from a lack of formality and poorly defined definitions. Although many machine learning interpretability methods and studies have been developed in academia and other institutions, they do not often form an integral part of machine-learning workflows or pipelines.

This chapter examines the role of explicable AI in clinical decision-support systems from technological, legal, and ethical perspectives.

There are many applications of XAI within the healthcare industry. The concept of explainability has many implications for all stakeholders. Developers, doctors,

# Chapter 7
# Explainable AI Driven Applications for Patient Care and Treatment

**Mukta Sharma, Amit Kumar Goel, and Priyank Singhal**

**Abstract** The continuous development of technology has saved countless lives and improved the quality of living. Artificial Intelligence is reshaping the healthcare industry from hospital care to clinical research, drug development, to insurance, and has been able to reduce costs and improve patient outcomes. Most AI system works as a black box with little or no explanation which results in a lack of trust and accountability among patients and doctors. This chapter is written with the intent to share with the audience how exquisitely the health care sector has integrated with the technology. The chapter initiates with a brief description of the use of Artificial intelligence and technology in the health domain, and how computers are helping not only doctors, but patients, health care departments, and Insurance companies. This chapter later focuses on various AI-driven Applications which are used for patient care and treatment. This chapter shed light on the purpose and benefits of XAI along with a few real examples.

**Keywords** Electronic health record · Artificial intelligence · Machine learning · Deep learning · Explainable artificial intelligence · Clinical decision support system

## 7.1 General

Society is getting enormous benefits from the advancement and innovation in technology. Technology has become an indispensable aspect of our lives. The continuous

M. Sharma (✉) · A. K. Goel
Delhi, India
e-mail: hod.bca@tips.edu.in

A. K. Goel
e-mail: amit.goel@galgotiasuniversity.edu.in

P. Singhal
Moradabad, India
e-mail: priyank.computers@tmu.ac.in

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_7

development in technology has saved countless lives and also have improved the quality of living. Technology is proving to be very relevant in the Health care sector, starting with EHR where the records of the patient are maintained electronically, a system to compile patient's medical history, which includes patients' past and present details. These days the data is collected through various devices like smartwatches, bands, patient's email, apps, etc. to provide a better monitoring system that will help in analyzing the health information.

Any IT gadgets or programming designed to enhance emergency clinic and authoritative efficiency, provide new bits of understanding about medications and therapies or improve the general nature of treatment is referred to as medical services innovation. Artificial Intelligence is being used extensively in various domains, like the reviews about the most-watched movie or series on Netflix, the most purchased product on Amazon, traffic congestion on road can be predicted by Google Maps, etc. are a few instances of the use of the AI.

AI technologies are reshaping the healthcare industry from hospital care to clinical research, drug development, to insurance, and have been able to reduce costs and improve patient outcomes. With the use of Artificial Intelligence in the subsequent years, it has been observed that there is a paradigm shift from what technology can do, to how technology should be used responsibly to improve health care services and patients' health. As most the AI system works as a black box; with very little or no explanation it results in a lack of trust and accountability amongst patients and doctors. The results generated by the AI tools could not be cross verified, and in case if they are generating a wrong decision; could lead to disaster specifically when it comes to the health care sector, which involves human life and one wrong decision can ruin a life. Therefore, it is essential and very crucial to use XAI, as it provides an explanation in natural language for a better understanding and rational decision making.

The present medical care industry is a $2 trillion behemoth (https://builtin.com/healthcare-technology). The apps and other health care devices with the use of AI have improved and helped in diagnosis, as it is more comfortable scanning and observing the pattern related to health like heart rate, blood pressure, footsteps took, calories burn/taken, to even monitor the sleep quality, etc. Artificial Intelligence is supporting emergency clinics by taking the decisions for better diagnosis based on analyzing and predicting the data.

With the enhancement in Artificial Intelligence, the medical domain has created robotic surgeries, where actually the physician is not even in the operation theater with the patient. The patient might be at the clinic or hospital in his home town eliminating any stress and hassle of traveling. Robotic surgeries also allow a minimally-invasive procedure that reduces the scars, and pain; which helps the patient cure fast (Bouronikous 2013). AI systems, such as deep learning or machine learning as the name suggests the machine is being trained by taking inputs and later producing outputs with no decipherable explanation or context.

Erik Birkeneder, a medical device, and digital health expert, in an interview with Forbes, "We can't be sure an AI system will discover those outliers or otherwise appropriately diagnose patients if it isn't properly trained with the relevant data and we don't understand how it makes its decisions" (https://www.capestart.com/resources/blog/how-explainable-ai-for-health-care-helps-build-user-trust/#:~:text=When%20and%20in%20what%20context,undetectable%20by%20the%20human%20eye).

Many of the AI algorithms are really insightful an algorithm to estimate the brain age is based on more than 5000 brain scans using a deep learning algorithm and is good in predicting the age and identifying if someone is getting cognitive decline or dementia and also has the capability to trace back the neural network and the changes in the brain due to the age or any other reason. Similarly, the genetic algorithm works very fine, so the algorithms refer to the image visualization and take the decisions based on the visuals, the results are often correct.

Explainable AI is the need of the hour, though XAI has been in the existence for approximately 40 years. It is gaining good popularity now as people are using Artificial intelligence extensively in almost all domains; especially in medicine where we are dealing with human lives we need to be assured, need to trust the solution/output given by AI system. The human need to understand why this decision has been taken or why this diagnosis has been proposed by the AI system. Designers and developers need the ability to explain to improve system robustness and allow diagnostics to avoid prejudice, injustice, and discrimination, as well as to raise user trust in why and how decisions are made. It is essential to give people a feeling that they can trust the software, the output should be interpretive (predicted or inferred), especially in cases where the images are synthesized. In short, XAI is required where the user needs an explanation to make a decision.

## 7.2 Benefits of Technology and AI in Healthcare Sector

From the invention of X-ray equipment to advances in surgical techniques, technology has improved our health and prolonged our lives (Hosny et al. 2018), Scherman (2019). Continued developments, and research in innovations that cure illnesses especially using Artificial Intelligence, training the devices to not only collect the data through sensors, and actuators but also to analyze the data using numerous algorithms which can predict with accuracy and precision (Mojsilovic 2019). Technology is helping us start by maintaining and keeping the patient's records handy through EHR (Electronic Health Record) instead of conventional paper-based manual methods. Also, the technology has made it possible to connect through Telemedicine, use remote monitoring health, and also share our health information through wearable and sensors technology with our doctors (Gulavani and Kulkarni 2010). Sequencing the human genome has been one of the greatest advancements in medical technology (McDonough 2021).

The innovation is certainly helping and permitting us to analyze a huge amount of data. The way Amazon has imagined Alexa, which is a virtual assistant based on an AI framework. Alexa helps in responding/answering conversational questions immediately, even in a noisy environment. Another AI application can help examine the information identified with individuals' wellbeing, permitting us to analyze the patterns they could turn into the key to well-being screening, early analysis, and treatment plan for a patient. The data accumulated by technology and sensors can have various advantages (Bouronikous 2013; Laal 2012; Luci 2015), for example:

- **Reduced healthcare cost and enhanced speed**—With the assistance of innovation, the data provided by the sensors can help in observing the well-being of the patients living at remote places. Real-time cautions can likewise assist a patient with counselling a specialist before health deteriorates. This remote monitoring of heath also eliminates hospital room expenses and staff costs. Splendid advancement can streamline claims planning, and cut costs by a colossal edge (Patel 2022).
- **Reducing healthcare waste**—According to the World Health Organization, healthcare waste accounts for about a quarter of all waste produced. An approximate 16 billion injections are issued per year around the world, however, the disposal of all needles and syringes is not done properly. Measures to guarantee the effective and ecologically sustainable disposal of medical wastes might assist to minimize harmful health and environmental implications. Providers and health insurers should follow these three ways to maximize healthcare costs during a period when 25% of spending is deemed unsustainable, (Thimbleby 2013), WHO/Unicef (2018).
- **Virtual Reality is helping in fighting Depression and Mental Health**—Approximately more than 800,000 peoples commit suicide every year; generally, because of emotional well-being issues. Psychological maladjustment can't be recognized by taking blood tests or breaking down information in an electronic clinical record. The data can suggest subtle traces of problems provided that they are well analyzed. Clinical psychologists and doctors identify the behavior and perform a cognitive test on the patients to know the situation. Especially the patients who have been born with any kind of prior traumas, doctors gradually trained the patient's brains to talk and build up tolerance through exposure treatment, until such memories no longer negatively affect them (Builtin).
- **New medicine and therapies are being developed**—According to a recent report, it takes at least 10 years for a drug to make the journey from discovery to the marketplace, at an estimated cost of $2.6 billion. The probability of a drug entering clinical trials being approved is currently estimated to be less than 12%. Incorporating AI into the drug-discovery process can have a significant impact on the development of safer and more successful drugs (Laal 2012).
- **AI-based Apps as a Scheduler**—To keep a human healthy, two important things need to be done, eat healthily and on time and follow a fitness regime. AI helps in analyzing the medical record and scheduling the fitness routine (Ksolves.com 2021).
- **AI for visually impaired or disabled people—**Many of the researches are going on to help people with any disability. AI-based devices are available in the market from gloves, shoes, etc. which will help and guide the disabled person (Ksolves.com 2021; Patel 2022).

- **AI for old people**—In research conducted in Japan observed that many clinics and old age homes have AI pets to keep the old people engrossed and stay connected with the devices. They have tried to use technology by designing AI robots in form of pets to give emotional support to the patients.

## 7.3 Most Common AI-Based Healthcare Applications

In recent times, AI has been benefiting the patients, doctors, and admin staff; without human intervention can complete task at a faster pace and are the talk of almost every conversation, especially in the medical domain. It has been observed that AI is often discussed by the modern medical industry to identify and diagnose the disease. There are numerous benefits of using artificial intelligence in today's contemporary medicine, but at the same time there are several issues that are bothering people; one such concern is a miss of the "human touch" in this people-oriented profession where people need to be supported emotionally as well and the trust; the patients have on the doctors mentally strengthen the patient and positive attitude heal the patient faster.

Artificial intelligence (AI) in modern medicine is used to describe the usage of AI software and pre-programmed processes to detect and treat patients that require medical treatment. Besides analysis and cure, there are a number of other processes that must be completed to appropriately take care of a patient designated, which may appear to be trivial tasks, including:

i. Gathering data from patient talks and examinations
ii. Dealing with and analyzing the results
iii. Obtaining precise identification by utilizing a multitude of data sources.
iv. Selecting an acceptable treatment method
v. Organizing and monitoring the treatment plan
vi. Observation by the patient
vii. Rehabilitation and continuing plans.

Healthcare has a wide range of applications, from identifying genetic code connections to powering surgical robots. Predictive, comprehending, reading, and acting machines are being reinvented. Artificial intelligence has been a benefit to the healthcare business in general.

Automation is delivering enormous benefits as creativity continues to flourish. Some applications that help to improve health care accuracy in addition to having a specialist solution that saves time and money in the treatment process (Datta et al. 2019; Nicholson 2019; Pawar et al. 2020).

- **Diagnostic Imaging Interpretation**—Deep learning programs and technology classification is used to provide AI-based imaging systems with algorithms that can read photos swiftly. Buoy Health, one of the most popular AI-based diagnostic checkers, uses an algorithm to help with sickness treatment (Your Team in India 2020). For instance, in the case of Lung cancer screening and to help detect pulmonary nodules, in many cases, early discovery can save a patient's life. Artificial intelligence (AI) can assist in recognizing and categorizing these nodules as benign or cancerous. Similarly, in the case of abdominal, mammography, brain tumor, and many more cases; artificial intelligence can interpret and evaluate whether they are benign or malignant. In the case of skin cancer, deep learning algorithms are handling and help detect suspicious areas (Hosny et al. 2018).

- **Accuracy**—According to the research diagnosis done by doctors are 71.40% accurate and diagnosis based on AI and ML is 72.52% accurate (Leibowitz 2020). Doctors are adopting a more contemporary strategy that emphasizes prevention and data collecting. This involves genetic data collection, wearable gadgets, and electronic healthcare system developments. Apple watches, Fitbits, Garmin watches, and other fitness trackers monitor your heart rate and activity levels (Luci 2015; Medttech; Your Team in India 2020).
- **Interactive Assistant for Fitness**—AI businesses have created digital health aides that focus on augmented reality, cognitive computing, speech, and body motions. A virtual health assistant is a one-of-a-kind method for reducing the number of trips to the hospital (Rauv 2017; Your Team in India 2020).
- **Bots that provide customer service—**Natural language processing (NLP) and sentiment analysis were used to construct collaborative Chatbots. Patients can ask inquiries regarding bill payment, appointments, and medication refills 24*7 all through the year (Rauv 2017; Xu et. al. 2019; Your Team in India 2020).
- **The Process of Robot-Assisted Surgery**—When acquiring insights, doctors might use pre-op medical data to acquire information. Specific movement, robotic arms, and magnetic imaging are all characteristics of the robotic surgical system (Ksolves.com 2021; Your Team in India 2020).
- **Digital Consultation**—There are numerous health care apps available, which will respond to and would handle the patients' queries in case the concerns raised by the patient need real doctors' intervention, and the call is transferred to the real practitioner (The Medical Futurist 2021). It uses emotional artificial intelligence to provide personalized medical consultations. These days the bots are the first line of primary care (Xu et. al. 2019).
- **Health Observation**—Artificial intelligence (AI) and relevant technologies such as ML and DL are used to monitor the patient's health. When modifications are made, the applications send notifications to the user.
- **Drug Creation**—The search for novel drugs is predicted to become faster, cheaper, and more effective as a result of machine learning and other technologies. When it comes to generating new drugs, clinical trials take a lot of time and money (Ksolves.com 2021; Xu et. al. 2019).
- **Machines that are linked together**—The unique artificial intelligence makes the operation both simpler and more intuitive (Xu et. al. 2019).
- **Electronic Health Records (EHR) Standard**—Traditionally, physicians would manually record or type findings and patient data, and no two were alike. Interactions with patients, clinical diagnoses, and future therapies can all be augmented and reported more reliably (Xu et al. 2019).

## 7.4 Issues/Concerns of Using AI in Health Care

According to a survey conducted by Accenture, "The AI healthcare market will expand at a compound annual growth rate of 40% by 2021 (Bresnick 2017). However, adoption in healthcare is still in its early stages." Here are some of the AI-related problems and concerns in healthcare.

- **Data Accessibility**: the training of AI systems necessitates a large quantity of data from many sources, including electronic health records. Data is frequently spread across several systems. Handle such a huge data and too fragmented data enhances the chances of inaccuracy, minimizes the database completeness, and also expands the cost of acquiring the data (Patel 2022).
- **Wounds and Fault**: One major concern is that if at any point in time, AI systems diagnoses or handle the patient incorrectly; it will be leading to a disaster or patient injury. If an AI system prescribes the wrong medicine, fails to discover a tumor, or assigns a hospital bed to the wrong patient, then the patient may suffer harm (Patel 2022).
- **Questions about privacy**: Developers are enticed to collect data from a high number of patients while working with enormous datasets. Some patients may be concerned that their privacy would be violated as a result of this data collection. As a result of data sharing between huge health systems and AI companies, lawsuits have been brought (Patel 2022).
- **Bias and inequality**: There is a risk of prejudice and inequity in healthcare AI. The AI system may be biased. As the machines learn from the data they're given and may incorporate biases based on that data (Patel 2022; Dilmegani 2017).
- **Safety and Transparency**: IBM Watson for Oncology has come under fire for allegedly making "unsafe and incorrect" cancer care guidelines. Instead of using actual patient data, the program was only trained with a few "synthetic" cancer cases (Price II 2019; Patel 2022).
- **Technical Debt**: In the last five to seven years, the new AI techniques based on deep neural networks have achieved incredible results. Few individuals possess the technical skills required to solve the full spectrum of difficulties relating to data and software engineering Limited data and changeable data quality will be a common problem for AI solutions (Price II 2019; Patel 2022).
- **Unexplainable AI Models**: In order to produce better performance, most AI models become more complicated. Both healthcare companies and patients are concerned about the lack of logic. To function appropriately.
- **Strict monitoring protocols to avoid diagnostic errors**: Diagnostic mistakes account for 60% of all medical errors, causing 40,000–80,000 fatalities per year (Kelly et. al. 2019; Price II 2019; Patel 2022).

## 7.5 Why Explainable AI?

Symbolic AI was the first form of AI, in which logic rules represented knowledge. There was no ability to learn and a weak ability to deal with ambiguity. Then there's Statistical AI, which uses huge data to train statistical algorithms for specific areas. There is no contextual capacity and only a bare minimum of explainability. As depicted in the figure below the results of AI creates confusion (Pawar et al. 2020; Turek 2016; Thimbleby 2013) (Fig. 7.1).

Later systems were constructed using Explanatory models (Ahmad 2020). Systems learn and reason with new tasks and situations. As shown in Fig. 7.2, how AI is different from XAI? In AI it just computes and gives end result; why this result has been given no explanation is provided. XAI will give explain the result, which will give better clarity and understanding of the decision. Explainable AI is a field where techniques are developed to clarify AI system predictions. In this chapter, XAI is explored as a strategy for using AI-based systems to analyze and diagnose health data. In the field of healthcare, accountability, outcome tracing, and model improvement are all essential. XAI can be used to achieve transparency in the healthcare industry. It is required to make it easier to share data about a patient's medical history with doctors and practitioners.

In recent years, AI researchers have worked to bring neural networks out of the shadows and make them more apparent. The initial AI Models built were based on a black box in which when the result is produced it is difficult to trace the detailed information about why this result has come. Let us understand with an example—if a company has implemented the AI for fraud detection so the machine will give a scoring or a rank but will not give a detailed explanation for the same. Whereas as depicted in Fig. 7.3, XAI explains the predictions made by the machine, which will help the organization understand with logic and clarity the predictions made by the machine. Similarly, if a customer's loan has been rejected, he can see the details and work on improving his score (Bizarro 2020). Similarly, if a person has a family

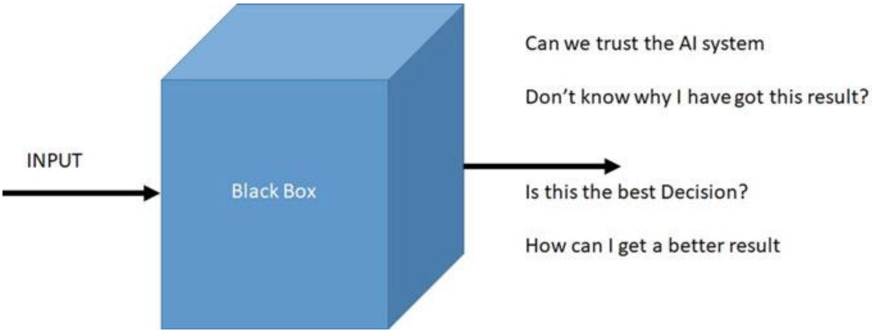

INPUT
Black Box
Can we trust the AI system
Don't know why I have got this result?
Is this the best Decision?
How can I get a better result

**Fig. 7.1** AI may create confusion as a decision is given without any explanation

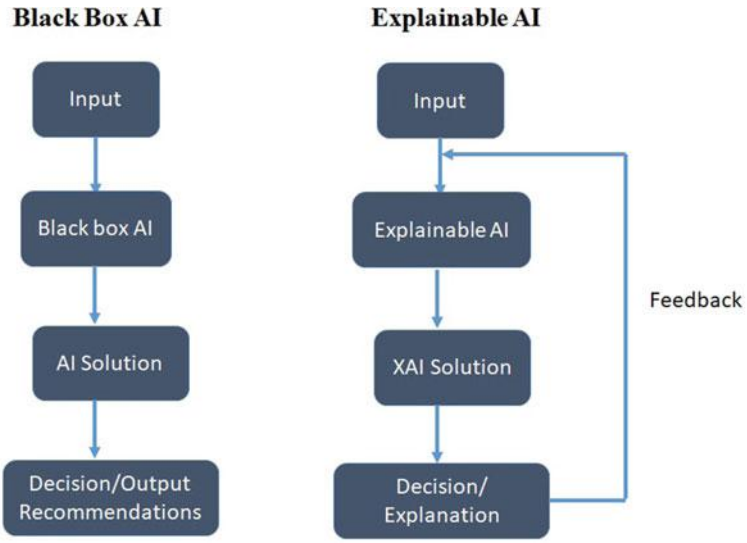
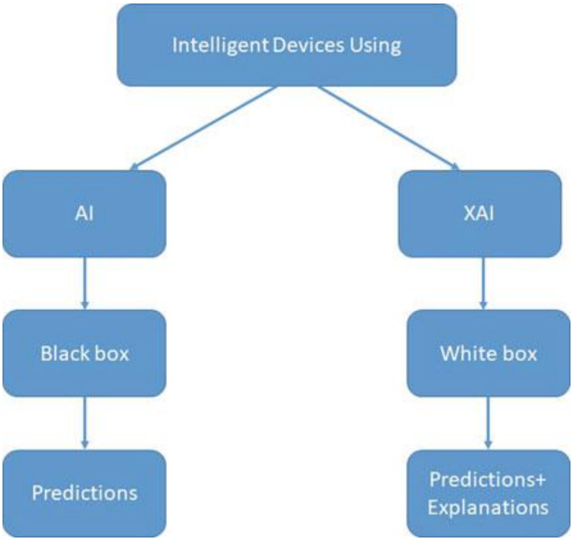


**Fig. 7.2** Need for XAI

history of any disease such as Diabetes or Cancer, can get the tests done and based on the analysis, and the results doctors can suggest a lifestyle, diet for prevention.

IBM has announced AI Explainability 360, describing it as "a robust open-source toolset of cutting-edge methods that aid machine learning model interpretability and explanation" (Mojsilovic 2019). AI must be viewed as black boxes, having internal inference procedures that are impenetrable to humans and undetectable to

**Fig. 7.3** XAI works as a white box that explains the results produced



the observer. The two most important aspects of XAI are transparency and post-hoc interpretation. In the eyes of developers, the transparency architecture shows how a model works, which includes feature relevance and evaluating & comparing the model.

Relevant features (specific data) for the study are provided like transaction amount, location, payment method, etc., and so on over a period of time can be used to detect fraud and prevent the same. The average money withdrawn from an account per month is an example of a feature. XAI assigns a value to each feature. Take a look at the features listed below:

- average monthly credit card charges in dollars.
- This card is likely to be used in this merchant category in this zip code.
- in the last week, the number of unique clients who utilized this IP address;

The model ranked the characteristic 'the average amount of credit card charges per month' in this case. Organizations can monitor and also can alter the priorities to improve the overall fraud detection and prevention. Later organizations can use model evaluation to assess the model's performance before releasing it. XAI model evaluation and comparison, is a critical component as organizations gain from testing the performance of a machine learning model since they can assess how accurate the model is and what the false positive rate is before relying on it for fraud detection and prevention. If the false positive rates do not satisfy the organization's objectives, the team tweaks the model until it performs optimally. Model comparison, as the name suggests, provides a side-by-side comparison of how different models perform when compared. Organizations can choose the best fraud detection models to deploy into production.

Let us shed a light on the components of Explainable AI, to attain transparency and post-hoc interpretation. Post-hoc extracts relationships between feature values and predictions, the behavior of a model. Model-agnostic methods can be used on any type of model, but model-specific methods can only be used on one type of model. They reject ante-hoc approaches, which build explainability into the model's structure, making it explainable even before the training phase is completed.

It attempts to (a) comprehend the model structure, such as decision tree construction; (b) understand single elements, such as a logistic regression limitation and (c) know the strategy for training, such as finding the solution in convex optimization. The post-doc interpretation explains why a result is presumed in the eyes of the customers. It tries to (d) provide analytic assertions, such as why a product is recommended on a shopping website; (e) provide visualizations, such as a saliency map for displaying pixel value in an object classification result; and (f) many transparent, interpretable algorithms such as K-nearest-neighbors, Convolutional Neural Networks, etc. are used to support the results (Dilmegani 2021; Xu et. al. 2019) (Fig. 7.4).

The Explainable AI (XAI) program aims to develop a set of machine learning techniques that will:

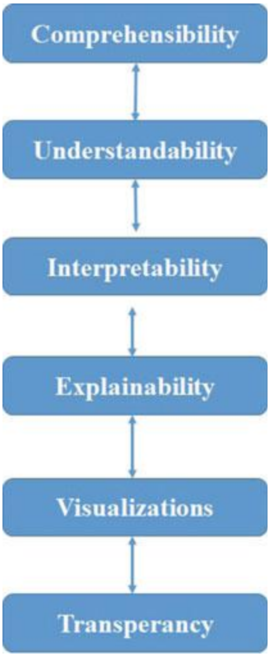


**Fig. 7.4** Components of explainable AI

- Enable users to comprehend, effectively manage and adequately trust, the emerging generation of artificially intelligent partners.
- Producing more explainable models while preserving high learning performance (prediction accuracy).

Various XAI solutions have been proposed in the recent past, and have also been applied to the healthcare sector. Some XAI models are self-explanatory, based on the decision sets, which have influenced largely the prediction of several diseases like diabetes, asthma, lung cancer, etc. by seeing the patient's health record. The patient's records are self-explanatory as they are developed by mapping an instance of data to an outcome using IF-THEN rules. For example, decision sets will learn to forecast lung cancer given the following circumstances.

**Predict lung cancer**: If the individual smokes and has a history of respiratory disease. Self-explainable AI models have the drawback of limiting the amount of AI models that can be used to increase accuracy. There has been a surging interest in XAI techniques that can explain any AI model to address explainability in a larger number of AI models.

Model-agnostic XAI procedures are those that are unaffected by the AI model that needs to be explained. One of the most extensively used model-agnostic approaches, ***Local Interpretable Model-Agnostic Explanation*** (LIME), was presented by researchers as a framework for explaining predictions by quantifying the contribution of all the components involved in calculating prediction.

Researchers utilized LIME to describe how Recurrent Neural Networks (RNNs) forecast heart failure, and their explanations helped them identify the most frequent health problems that raise the risk of heart failure in people, such as renal failure, anemia, and diabetes. In the healthcare arena, various model-independent XAI approaches like Anchors and Shapley values have been developed and are actively used. An outline/ framework was specified for using human reasoning expertise in the development of XAI approaches, to improve details by incorporating the user's cognitive skills. The methodology developed could be used for any specific fields, such as to improve healthcare, and to provide user-friendly comprehension of how AI-based systems that apply XAI techniques at various phases enhances the clinical decision-making work.

There are certain challenges in putting XAI techniques into practice. XAI has created explanations to benefit the end-users, who might be physicians with medical domain knowledge or ordinary people. It is possible to create proper user interfaces for successfully displaying explanations.

## 7.6 History of XAI

Before talking about the XAI history, let us have a quick glimpse of AI. AI is where the machines depict and mimic human intelligence. Like—Self-driving cars, games using AI like chess, Amazon echo, Alexa, Siri, Google Alpha Go, IBM Watson, chatbots on websites working as virtual assistants (Kalinin 2020), etc. One can see many sci-fi movies like The Terminator, Star Trek, ex Machina, etc. AI is also used by E-Commerce organizations to suggest products based on previous purchasing patterns. Pepper recognizes human faces with a few emotions, Da Vinci Surgical System can perform minimally invasive surgeries, and Google Duplex can make reservations over the phone.

AI has moved from making intelligent machines to Machine learning, which has the ability to learn without being explicitly programmed. ML consists of 3 techniques Supervised learning (An input is mapped with an output, data sets are mapped—help to predict the next value), Unsupervised (Data-driven approach and clusters are formed), and Reinforced learning (learn from Mistakes—like these days gaming apps are based on this). It uses data to detect patterns and adjust accordingly, developing programs that can teach themselves to change and grow when needed and enables computers to find hidden insights using algorithms. In short, ML automates analytical model building. Numerous ML Techniques are available like Classification, Categorization, Clustering, Trend Analysis, Anomaly Detection, Visualization, and Decision Making. ML is used in Image processing, health care, robotics, text analysis, video games, and data mining. ML is used in various applications like—spam filtering, information extracting, and sentiment analysis. ML is a subset of AI and superset of DL. ML models need human intervention to reach the optimal outcome.

Deep Learning makes predictions independent of human intervention. DL makes the computation of multi-layer neural networks, based on the human brain. A few examples of DL, ImageNet a database of 14 million labeled images used to train neural nets, 2012, Google Brain team trained the neural networks by watching unlabeled images of cats from frames of YouTube Videos. In the year 2014, Facebook's Deep face was released identifying the face with 97.35% accuracy with a training set of 4 million images. Alpha Go developed by Google Deep mind defeated the 18-time world champion Lee Sedol in the year 2015. Deep learning is reshaping healthcare through image analytics and diagnosis and drug discovery and precision medicine.

Explainability in XAI derives from a combination of strategies that improve machine learning models' contextual flexibility and interpretability. There is no formal definition, however, it can be defined as the ability to draw conclusions based on conceptions (as a person) rather than probability alone. It can be seen as "contextual reasoning". An explainable Artificial Intelligence generates information or justifications to make its operation understandable or simple to comprehend. The next generation of artificially intelligent partners, with the help of XAI, will develop a set of machine learning algorithms to assist people in comprehending, trusting, and managing the diagnosis.

Explainable AI isn't a new notion. In the first work on explainable AI, which was published in the literature forty years ago, certain expert systems used rules to explain their conclusions. Since the dawn of AI, experts have advocated that intelligent systems should be used to explain AI findings, particularly when it comes to judgments. If a rule-based expert system refuses to accept a credit card charge, it must provide an explanation. The principles and knowledge of expert systems are simple to comprehend and infer since they are explained and established by human experts. A logically structured decision tree is a common strategy as shown in the following figure to demonstrate that why a loan application gets rejected on what parameters using decision tree (Fig. 7.5).

Similarly, XAI, helps in health care, let us see with the following figure, which helps the patient knows about the lump detected is benign or cancerous (Fig. 7.6).

## 7.7 Explainable AI's Benefits in Healthcare

Health care workers utilize AI to speed up and improve a variety of functions, including forecasting, risk management, decision-making, and even diagnosis, by scanning medical pictures for anomalies and patterns that are undetected to the naked eye. Many health care practitioners now use AI as a critical tool, but it is often difficult to understand, causing dissatisfaction among clinicians and patients, especially when making high-stakes decisions.

Machine learning's lack of explainability limits its use in healthcare applications where decision-makers need to understand the underlying reasoning. If artificial intelligence (AI) is unable to justify itself in the field of business, it will not be implemented on a big scale. If anyone is in charge of healthcare, the risk of taking

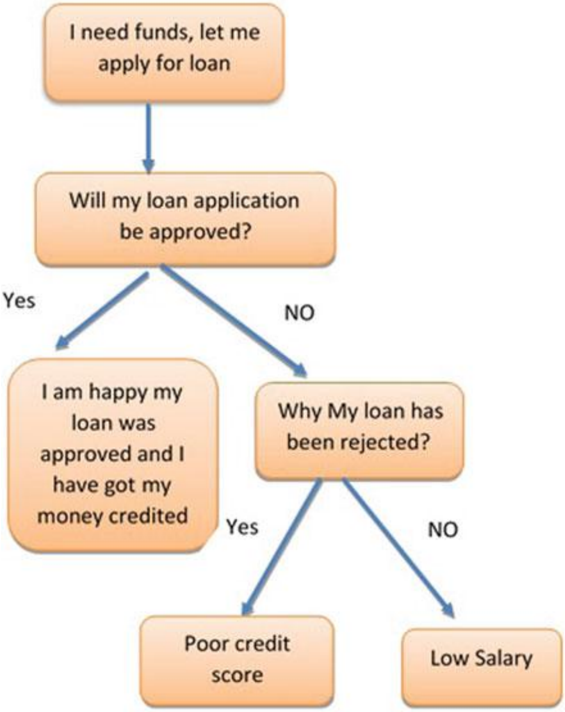

I need funds, let me apply for loan
Will my loan application be approved?
Yes
NO
I am happy my loan was approved and I have got my money credited
Why My loan has been rejected?
Yes
NO
Poor credit score
Low Salary

**Fig. 7.5** Illustrates an example of a decision tree, which is constructed by working down from the top, level by level, according to the reasoning stated

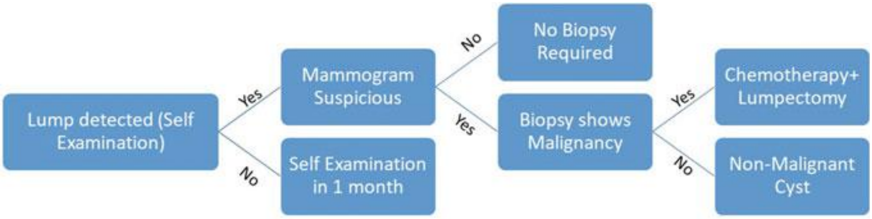


**Fig. 7.6** Decision tree for detecting breast cancer

a poor/wrong decision may outweigh the benefits of precision, pace, and decision-making efficiency. This will, harm the environment. As a result, its reach and utility will be severely limited. As a result, it's important to take a close look at these issues. Until a model can be implemented in the healthcare domain, standard tools must be developed. Explainability is one such method (Gulavani and Kulkarni 2010).

XAI increases medical practitioners' and AI researchers' confidence in AI systems, resulting in the more widespread use of AI in healthcare. The heart disease dataset as an example of how explainability approaches can be used to build trustworthiness when using deep learning systems in healthcare. The adoption of XAI comes with a slew of advantages. No matter what industry you work in, these advantages will help you and your company succeed. These are only a few of the main benefits of using XAI, according to Philip Pilger Storfer, a Quantum Black Data Scientist and XAI specialist:

- Developing user trust,
- Complying with legal obligations,
- Providing ethical reasoning, and
- Obtaining actionable and strong insights (Gill 2001).

Many companies are implementing XAI methodologies and techniques to achieve greater effectiveness, as shown by the findings described above. Explainable AI alleviates AI's problems and offers the following advantages (Gill 2001):

- **Trust and confidence**: Due to the uncertain existence of AI systems, gaining trust and confidence in doctors and patients is difficult. Users ask for answers from computers. In the pursuit of better efficiency, modern machine learning architectures are becoming more complex, often relying on black box-style architectures that provide computational benefits at the cost of model intelligence.
- **Detect and Remove Prejudice**: Since the system lacks clarity, users are unable to identify the system's flaws and biases. As a consequence, identifying and removing bias, as well as offering bias defense, becomes difficult.
- **Model Performance**: Model users are unable to monitor the model's actions due to a lack of knowledge.
- **Regulatory Standards**: Consumers are unable to assess whether or not the device complies with regulatory standards. Otherwise, the device could be affected.
- **Risk and Vulnerability**: It's important to be able to explain how systems deal with threats. Especially in circumstances where the user is unsure of the surroundings. Explainable AI assists in identifying it in a timely manner and taking effective action. But what if the device doesn't tell the consumer how to stop these dangers?

Explainable AI has accelerated the use of AI systems in healthcare. It is difficult for a person to make decisions because AI systems understand trends and make decisions based on Big Data. Explainable AI provides the following features (Gill 2001):

1. **Transparency**: The most important principle of Explainable AI is transparency. It is the algorithm, model, and features that the user can comprehend. Different users can need different levels of transparency. It is including appropriate explanations for appropriate users.
2. **Reliability**: The device offers reliable details. It should be in line with the model's production.
3. **Domain meaning**: The framework offers a user-friendly description that makes sense in the context of the domain. It is elucidating in the proper sense.

4. **Consistency**: For all forecasts, the interpretation should be consistent because different explanations will confuse the consumer.
5. **Generalizability**: The device should be able to explain things in a broad sense. However, it should not be too broad.
6. **Simplicity**: The system's description should be clear. It needs to be as transparent as possible.
7. **Reasonable**: It achieves the objective of each AI system's result.
8. **Traceable**: Explainable AI can track data and logic. The contribution of data in the production is revealed to the users. The user can track and solve logic and data problems.

## 7.8 XAI Has Proposed Applications for Patient Treatment and Care

Algorithms "that are inherently explainable" are the simplest method to develop functioning XAI in health care. Simpler solutions like decision trees, regression models, bayesian classifiers, and other transparent algorithms can be employed instead of sophisticated deep learning or ensemble approaches like random forests "without sacrificing too much performance or accuracy." XAI has been benefiting the medical practitioners and experts

1. **Assisted or automated diagnosis and prescription**: Chatbots can assist patients in self-diagnosis as well as doctors in diagnosis. Based on the symptoms identified by the patient, many healthcare organizations provide useful health and triage information. They do, however, note that no diagnosis has been made. This is to reduce their legal liability, but if the accuracy of chat bots increases, we may see chatbots delivering diagnoses in the future (https://medicalfuturist.com/top-12-health-chatbots/) Kalinin 2020; Kaushal et al. 2019).
2. **Prescription auditing**: Prescription auditing systems that use artificial intelligence (AI) can assist decrease prescription mistakes.
3. **Pregnancy Management**: Keep an eye on both the mother and the fetus to assuage the mother's fears and enable an early diagnosis.
4. **Real-time case prioritization and triage**: Prescriptive analytics on patient data enables real-time case prioritization and triage.

   - Jvion: The Cognitive Clinical Success Machine accurately predicts danger and comprehensively, providing prescribed actions that enhance outcomes.
   - Well frame: It flips the script by offering interactive care services to patients via their mobile devices. The Care Team's portfolio of clinical modules, which are based on evidence-based care, allows it to give a tailored experience.
   - Enlitic: Patient triaging solutions search incoming cases for various clinical findings, priorities them, and route them to the network's most suitable doctor.

5. **Personalized medications and care**: Assess the most appropriate treatment options based on patient data, decreasing costs and increasing care efficacy.

- GNS Healthcare: The business uses machine learning to match patients with the most effective treatments.
- Oncora Medicals: Software that helps health systems structure, interpret, and learn from their data in order to provide personalized care.

6. **Patient Data Analytics**: Analyze data from patients and/or third parties to uncover information and make recommendations. The organization (hospital, etc.) may use AI to analyze clinical data and derive deep insights into the health of patients. It allows for lower healthcare costs, more effective resource usage, and easier population health management (Daley 2018).

   - Zakipoint Health: The organization uses a dashboard to show all related healthcare data at the member level, allowing users to better understand risk and cost, as well as have personalized services and increase patient engagement.

7. **Surgical robots**: AI and collaborative robots are combined in robot-assisted surgeries. These robots are suitable for procedures that involve the same, repetitive movements because they can operate without being fatigued. AI can detect trends in surgical procedures, helping surgeons to improve their best practices and surgical robot control accuracy to sub-millimetre precision.
8. **Early diagnosis**: Analyze lab results and other diagnostic data to achieve an early diagnosis of chronic diseases.

   - Ezra: Ezra uses artificial intelligence to interpret full-body MRI scans and assist clinicians in cancer detection.

9. **Medical imaging insights**: Advanced medical imaging may be utilized to analyze and manipulate pictures, as well as to simulate future situations.

   - SkinVision: By taking pictures of your skin with your phone and going to a doctor at the right time, you can spot skin cancer early.
   - Medical imaging powered by artificial intelligence is also frequently used to diagnose COVID-19 cases and classify patients that need ventilator support. For example, Huiying Medical, a Chinese company, has developed a 96 percent accurate AI-powered medical imaging solution.

## 7.9 Future Prospects of XAI in Medical Care

The real healthcare advantage in the future would most likely come from the synergies gained by integrating the power of XAI-related technologies across the entire patient journey. Like, currently wearable devices could monitor heart rates, calories taken, sleep patterns, exercise levels, over time, blood glucose levels, and many more. In the future, this data could be synced to a central monitoring system that uses machine learning to detect irregular or undesirable behavior. The monitoring system will automatically alert the patient's physician and advise the patient to make an appointment (Fig. 7.7).

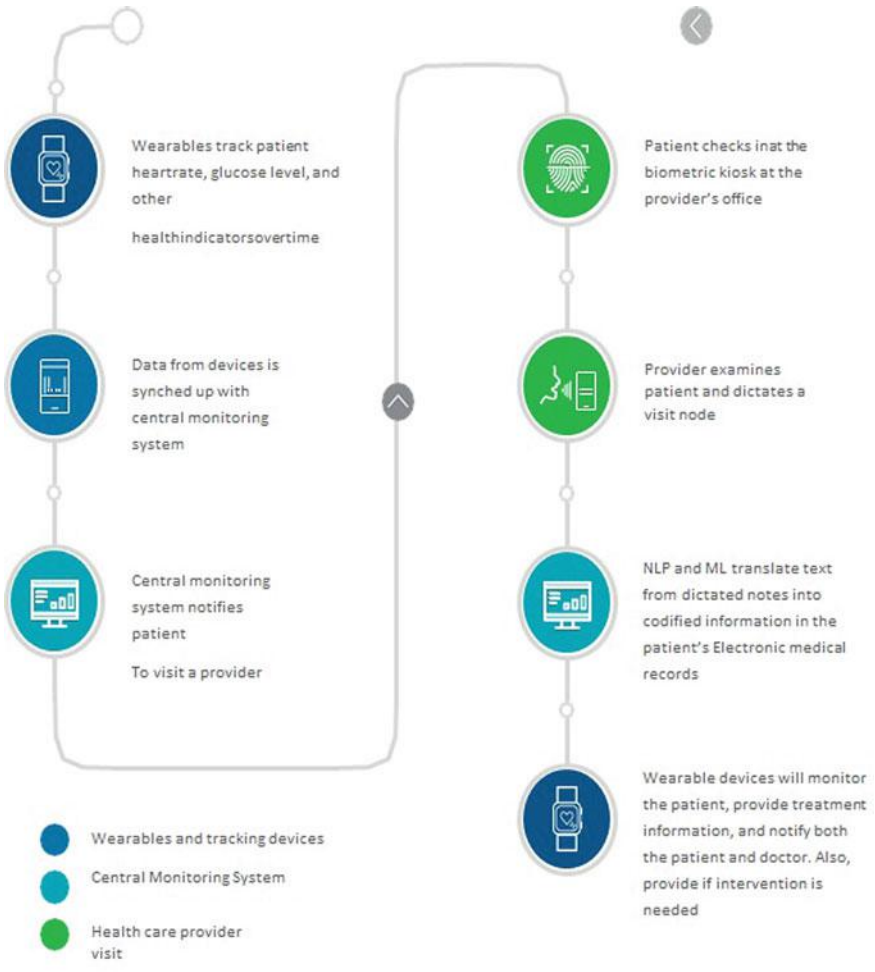

Wearables track patient heartrate, glucose level, and other
healthindicatorsovertime
Data from devices is synched up with central monitoring system
Central monitoring system notifies patient
To visit a provider
Patient checks inat the biometric kiosk at the provider's office
Provider examines patient and dictates a visit node
NLP and ML translate text from dictated notes into codified information in the patient's Electronic medical records
Wearable devices will monitor the patient, provide treatment information, and notify both the patient and doctor. Also, provide if intervention is needed
Wearables and tracking devices
Central Monitoring System
Health care provider visit

**Fig. 7.7** XAI in health care sector

## 7.10 Case Study on Explainable AI

Explainable AI can help give local or global explanations for single predictions. There are techniques like model agnostic, and model specific. Various algorithms produce non-explainable predictions like the random forest, SVM, NN, etc. There are methods like Intrinsic/Ante-hoc and Post hoc methods. The ante-hoc methods are transparent, can be said that they are based on a white-box approach. The algorithms like Decision Trees, KNN, Fuzzy, Bayesian, etc. are used for transparent and fair predictions. Post-hoc is based on algorithms like PCA, CAM, LIME (**L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations), or SHAP (**Sh**apley **A**dditive ex**p**lanations) (Daley 2018).

Diabetes is a very common disease in modern days due to lifestyle and eating habits. Diabetes is a condition in which the patient's blood glucose, often known as sugar, is usually either very high or too low. Glucose is the main source of energy that one gets from the food one eats. Although diabetes has no permanent cure, one can try to control/manage the sugar levels. It is necessary to keep track of the body. Explainable AI explains the data used for the prediction, their correlation, and EDA (Exploratory Data Analysis) to understand the hidden data patterns.

## 7.11 Framework for Explainable AI

XAI aims to make the model work transparently. It aims to select the right data and preprocess it for the model. It is required to have an accuracy of the model to predict the correct result, we need to be doubly sure when we are dealing with the health care domain as that is impacting the life and health of any individual. A single wrong result can have a threat to human life.

- Which feature influences more diabetes?
- What amount of Glucose do I need to maintain?
- Why did the system say that I can have diabetes in the future?

**The explainable nature of AI can help doctors**

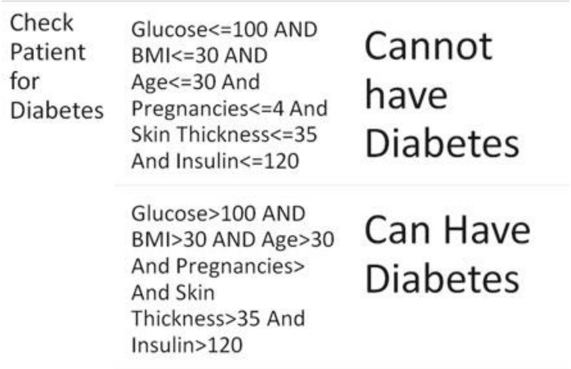
The Glucose value of a person influences the result more while predicting whether a person can have diabetes or not. Doctors and health practitioners can answer what Glucose value an individual need to maintain to have diabetes. The explainable nature of AI can help us look at the change of having diabetes in the future. LIME and SHAP, two of the most common feature-based techniques, are very similar in intent but take very diverse methods.

Unlike LIME, which can only provide local explanations, SHAP can provide both global and local explanations. The dataset is used to generate many plots that illustrate the dataset globally and provide details about the relationships between the features and their significance.

Explainability is a key to producing a transparent, proficient, and accurate AI system. It makes it easy for the enduser to understand the AI systems' complex work. (Fig. 7.8).

**Case Study's Conclusion**

Explainable AI's contribution to the Diabetes Prediction framework simplifies the intricate workings of AI systems for the end-user. It offers the user a human-centered GUI. Explainability is essential for creating a transparent, competent, and reliable AI system that can assist healthcare practitioners, patients, and researchers in comprehending and using the system.



| | | |
|---|---|---|
| Check Patient for Diabetes | Glucose<=100 AND BMI<=30 AND Age<=30 And Pregnancies<=4 And Skin Thickness<=35 And Insulin<=120 | Cannot have Diabetes |
| | Glucose>100 AND BMI>30 AND Age>30 And Pregnancies> And Skin Thickness>35 And Insulin>120 | Can Have Diabetes |

**Fig. 7.8** XAI in detecting diabetes

## 7.12 Conclusion

The disruptive impact of technology on the healthcare business is undeniable. Even though it is a sector that necessitates highly skilled employees with several years of education, it also necessitates a significant amount of infrastructure and instruments. The rise in global life expectancy and the ageing of societies have fueled a surge in healthcare innovation and technology. With the environment changing every year, it looks like field innovation is highly powerful.

The majority of AI systems are not accountable for their outcomes, which can often damage society or users by producing incorrect results. Explainable AI and its values improve the way a system operates by describing the algorithm, model, and features it employs. The domain of XAI must be defined and deployed in AI-based health systems continuously. Most AI system is not answerable for their result, which can sometimes harm society or the user. Explainable AI and its principle bring a change in the system's traditional functioning.

The value of improving explainable AI capabilities has begun to be recognized by the markets. Many companies like Microsoft Azure and Google Cloud Platform have influenced many AI adoption patterns.

## References

(n.a).: Top 12 benefits of AI in healthcare in 2021. Ksolves Emerging Ahead Always. https://www.ksolves.com/blog/artificial-intelligence/top-12-benefits-of-ai-in-healthcare-in-2021 (2021)

(n.a).: The top 12 health Chatbots. The Medical Futurist. https://medicalfuturist.com/top-12-health-chatbots/ (2021)

(n.a).: How explainable AI (XAI) for health care helps build user trust—even during life-and-death decisions. Capestart. https://www.capestart.com/resources/blog/how-explainable-ai-for-health-care-helps-build-user-trust/#:~:text=When%20and%20in%20what%20context,undetectable%20by%20the%20human%20eye