XAI in Healthcare: Black Box to Interpretable Models

Dr. Shikha Verma[1], Ms.Sandhya Avasthi[2], Ms.Priya Mishra[1], Ms. Meghna Gupta[1], Ms.Gunjan Agarwal[1]

[1]MCA Department, ABES Engineering College, Ghaziabad, India

shikha.verma@abes.ac.in, priya.mishra@abes.ac.in, meghna.gupta@abes.ac.in, gunjan.agarwal.ac.in,

[2]Department of CSE, ABES Engineering College, Ghaziabad, India

sandhya.avashthi@abes.ac.in,

Abstract

Explainable Artificial Intelligence (XAI) is a critical aspect of AI development, particularly in sensitive and high-stakes applications. Its benefits extend beyond technical aspects to encompass ethical considerations, user acceptance, and the responsible deployment of AI technologies. It refers to the techniques and methods used to make AI models and their decisions more transparent and interpretable to humans. Traditional machine learning algorithms, such as deep learning neural networks, are often considered "black boxes" because they lack transparency, making it challenging to understand how they arrive at specific conclusions or predictions. XAI aims to address this limitation by providing human-readable explanations for AI model outputs. There are many benefits of XAI including Transparency, Trust, and Confidence, Identifying Biases, Debugging and Improvements, Regulatory Compliance, Human-AI Collaboration, Education and Understanding, Insights, and Research.

XAI can help healthcare professionals understand how an AI system arrives at a specific diagnosis or treatment recommendation it helps in Enhanced Diagnosis and Treatment, Improved Patient Outcomes, Safety and Risk Assessment, Ethical AI Use, Regulatory Compliance, Patient Trust and Acceptance, Enhanced Diagnosis and Treatment, Improved Patient Outcomes, Safety and Risk Assessment, Ethical AI Use, Regulatory Compliance, Patient Trust and Acceptance, Research and Knowledge Discovery, Continuous Learning and Improvement. XAI is crucial in bridging the gap between complex AI algorithms and human understanding. It empowers healthcare professionals to harness the potential of AI while making informed and responsible decisions for the benefit of patients and the healthcare industry as a whole.

Keywords: Explainable Artificial Intelligence, NLP, Machine Learning, Explainable Model

1. Introduction

Applications of AI make it possible for autonomous systems to sense their environment, educate themselves, and come to their own conclusions. The aforementioned systems are specifically designed to handle extensive collections of data and employ advanced technologies such as machine learning and natural language processing (NLP) to yield outcomes. Nevertheless, a significant quandary arises from the fact that existing systems fail to provide us with the underlying rationale behind their decision-making processes. The provided output is presented without accompanying elucidation regarding the underlying causative factors responsible for its generation. It is imperative to acknowledge that within this context, there exists a discernible prediction, albeit lacking in accompanying justification.

The current situation can be characterised as a "black box" scenario, wherein the contents and mechanisms of the box are unknown to observers. Hence, it is plausible that this enigmatic "black box" possesses the potential to confine the utilisation and scope of artificial intelligence (AI). If artificial intelligence (AI) machines are incapable of engaging in transparent discourse with humans, their potential for achieving true intelligence will remain unattainable. Explainable artificial intelligence (AI) serves the purpose of enhancing the transparency of the "black box" phenomenon, thereby facilitating the ability of practitioners to engage in comprehensive analysis and comprehension of AI systems. The recent development represents a significant advancement toward enhancing the ethical framework of artificial intelligence.

Explainable AI (XAI) is an artificial intelligence program that explains the thought process behind reaching a particular result. Introducing a novel learning process that not only provides accurate predictions but also offers comprehensive explanations for the rationale behind each prediction.[1]. In the construction of an Explainable AI model, an additional layer is incorporated into the machine learning (ML) design to enhance its explanatory capabilities as shown in Figure1.

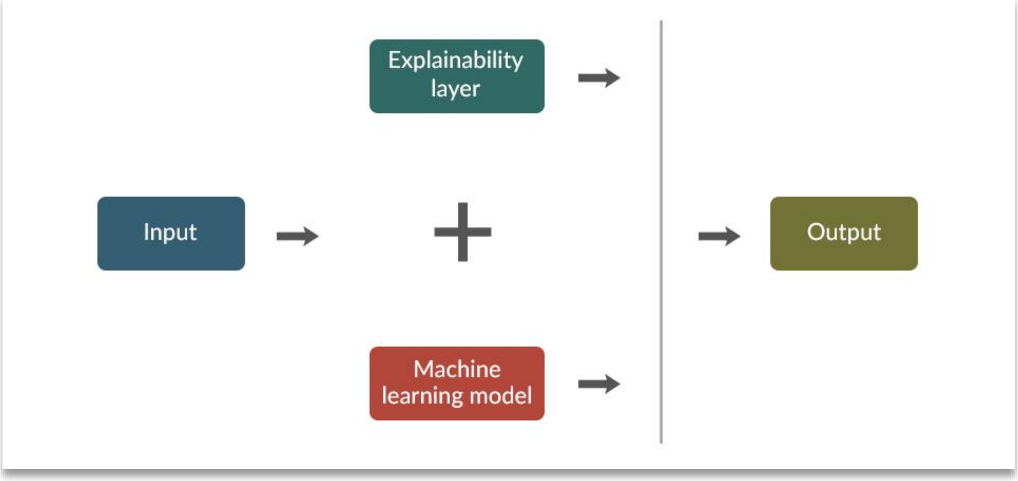

Figure 1. Explainable AI Model

The inclusion of an additional layer is of utmost importance due to its pivotal role in the enhancement of decision-making objectivity through the facilitation of explainability [2]. The utilisation of this technique facilitates the process of discerning and mitigating any inherent biases that may be present within the datasets.

The concept of explainability directs our focus towards various factors that have the potential to influence the outcome of a prediction. Explainable AI (XAI), alternatively referred to as Interpretable AI, is a burgeoning domain that directs its attention towards a diverse array of methodologies aimed at mitigating the inherent opacity of Machine Learning models, thereby engendering elucidations that resonate with human-level comprehension. Explainable artificial intelligence (XAI) refers to a comprehensive array of systematic procedures and methodologies that facilitate the comprehension and establishment of trust among human users in relation to the outcomes and outputs generated by machine learning algorithms. The primary objective of the Explainable AI (XAI) programme is to develop a comprehensive set of machine learning methodologies that fulfil two key objectives. Firstly, these methodologies should generate models that are more transparent and interpretable, without compromising their ability to achieve high levels of learning performance, specifically in terms of prediction accuracy. Secondly, they should empower human users to comprehend, place appropriate trust in, and proficiently oversee the forthcoming cohort of artificially intelligent collaborators.[2]

Future iterations of machine-learning systems will possess the remarkable capability to elucidate the underlying reasoning behind their decisions, delineate their inherent advantages and limitations, and effectively communicate their projected behaviour in forthcoming scenarios. The proposed approach to attaining the aforementioned objective involves the cultivation of novel or adapted machine learning methodologies, which will yield models that are endowed with enhanced interpretability[3] .The integration of these models will be complemented by cutting-edge human-computer interface methodologies, which possess the ability to convert the models into comprehensible and valuable explanation dialogues for the end user.

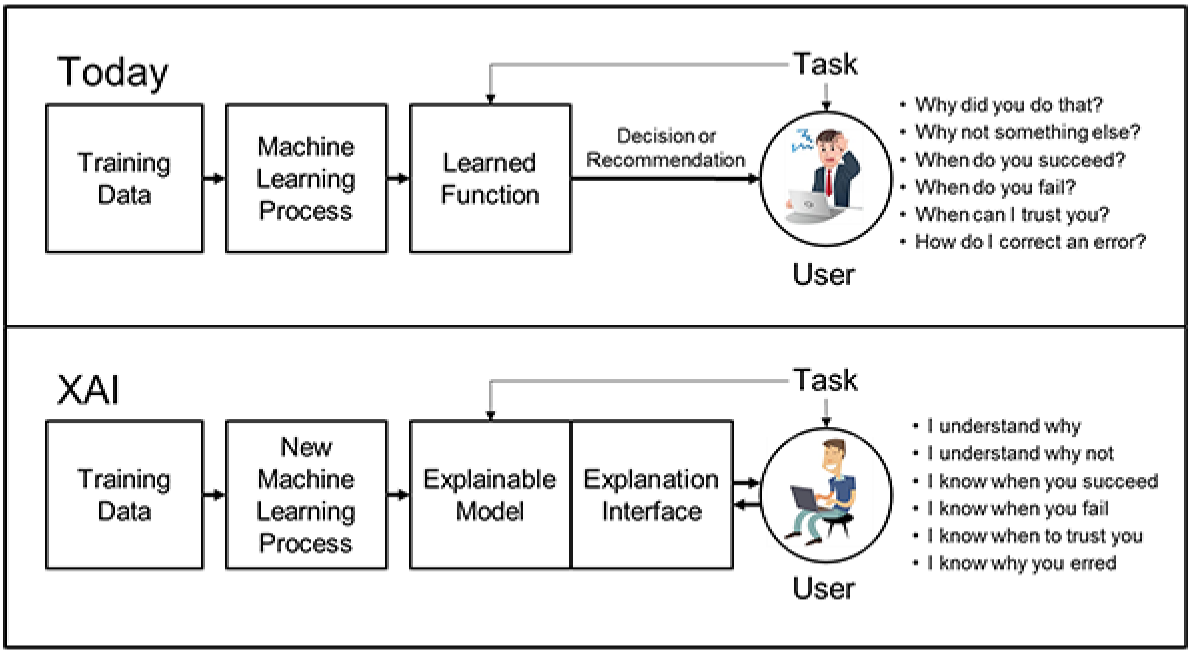

Figure 2. Comparison of XAI with Traditional Model

1.1 Significance of XAI

The significance of Explainable Artificial Intelligence (XAI) cannot be overstated, as it addresses a critical limitation of traditional black-box machine learning models. By providing interpretability and transparency, XAI offers a multitude of benefits that have far-reaching implications for various domains. Firstly, XAI enhances trust and accountability in AI systems. By elucidating the Explainable Artificial Intelligence (XAI) presents a range of benefits that encompass transparency, accountability, trust, and compliance within decision-making procedures. Additionally, XAI aids in the identification and mitigation of biases that may otherwise result in inequitable decision making. Through the provision of lucid and all-encompassing elucidations pertaining to the intricate process of decision-making, Explainable Artificial Intelligence (XAI) possesses the potential to facilitate the formulation of well-informed and resolute judgements by decision-makers [3]. Given the prevailing regulatory landscape, wherein industries and organisations are compelled to adhere to transparent and interpretable decision-making protocols, the integration of Explainable Artificial Intelligence (XAI) emerges as a viable solution to fulfil these mandates. Hence, explainable artificial intelligence (XAI) assumes a pivotal position in facilitating the attainment of organisational objectives and ensuring a competitive edge in a rapidly changing and dynamic landscape.

The acceleration and extensive deployment of artificial intelligence systems are directly proportional to the level of confidence instilled in them. By strategically positioning your business, you will enhance its capacity to cultivate innovation and gain a competitive edge in the realm of developing and embracing cutting-edge capabilities of the next generation.

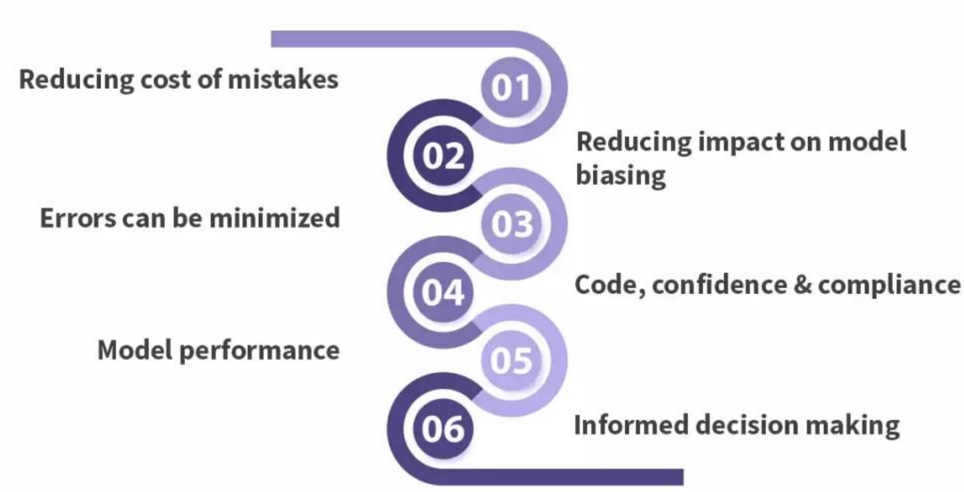
Reducing cost of mistakes
01
02
Reducing impact on model biasing
Errors can be minimized
03
04
Code, confidence & compliance
Model performance
05
06
Informed decision making

Figure 3. XAI Benefits

- **Reducing cost of mistake**- Erroneous prognostications exert a substantial influence on decision-centric sectors such as healthcare, finance, jurisprudence, and the like. The act of monitoring the outcomes serves to mitigate the potential ramifications of erroneous discoveries and facilitates the identification of the fundamental factors, thereby enhancing the foundational framework.[4] Consequently, the advent of advanced artificial intelligence (AI) technologies has engendered a notable enhancement in the credibility and authenticity of AI writers, rendering them increasingly reliable and practical for utilisation.
- **Reducing Impact of Model biasing-** The empirical findings have indicated a substantial presence of bias within AI models. Instances of bias have been observed in various domains, such as the presence of gender bias in the Apple Cards system, the manifestation of racial bias in autonomous vehicles, and the identification of gender and racial bias in Amazon Recognition. An explainable system possesses the potential to mitigate the deleterious consequences stemming from biased predictions by elucidating the underlying decision-making criteria.
- **Errors can be minimized**- It is a well-established fact that AI models inherently possess a certain degree of fallibility in their predictive capabilities. However, the integration of a human agent who possesses the qualities of responsibility and accountability can significantly enhance the overall efficiency of the system.

- **Code Confidence and Compliance-** Each instance of inference, accompanied by its corresponding explanation, has the propensity to augment the level of confidence within the system.[4] Certain user-critical systems, including but not limited to autonomous vehicles, medical diagnosis tools, and the finance sector, necessitate a heightened level of code confidence from users in order to achieve optimal utilisation. In light of the escalating demands imposed by regulatory bodies, enterprises find themselves compelled to swiftly embrace and operationalize Explainable Artificial Intelligence (XAI) in order to ensure compliance with governing authorities.
- **Performance evaluation** -In order to optimise performance, it is imperative to possess a comprehensive comprehension of the potential vulnerabilities that may impede progress. Enhancing the comprehension of model behaviour and the underlying causes of occasional failures facilitates the process of refining said models. The concept of explainability holds significant potential in the realm of artificial intelligence research[4]. By enabling the detection of flaws within models and biases inherent in the data, explainability serves as a potent tool for fostering trust among all users. The utilisation of this approach facilitates the validation of prognostications, the enhancement of computational models, and the acquisition of novel perspectives pertaining to the issue under consideration. The identification of biases within the model or dataset is facilitated by a comprehensive comprehension of the model's underlying mechanisms and the rationale behind its predictive outcomes.
- **Informed Decision Making**-Machine learning applications have emerged as a pivotal tool in the realm of business, primarily serving the purpose of facilitating automated decision making processes. However, it is frequently desirable to employ models primarily for the purpose of extracting analytical insights. As an illustration, one could employ a model to forecast sales within an expansive retail conglomerate by leveraging a comprehensive dataset encompassing variables such as geographical coordinates, operational hours, meteorological conditions, temporal context, product assortment, and physical outlet dimensions, among others. The proposed model exhibits the capability to prognosticate sales across multiple retail establishments, encompassing a comprehensive temporal scope spanning every day of the calendar year, while accommodating diverse meteorological circumstances. By constructing a comprehensible model, one can discern the primary catalysts behind sales and leverage this knowledge to enhance revenue generation.

2 Motivation of Applying XAI In Health Care

The prevailing sentiment is one of profound optimism regarding the potential of artificial intelligence (AI) to yield significant enhancements across the entire spectrum of healthcare, encompassing diagnostics as well as treatment modalities. The prevailing consensus posits that artificial intelligence (AI) tools are poised to serve as facilitators and augmenters of human labour, rather than supplanting the roles of physicians and other healthcare personnel outright. Artificial intelligence (AI) has reached a level of maturity where it can effectively assist healthcare professionals across a wide range of responsibilities. These encompass not only administrative tasks and workflow management but also extend to critical areas like clinical documentation, patient outreach, and even specialised domains such as image analysis, medical device automation, and patient monitoring. The integration of AI into healthcare systems holds immense potential for enhancing the efficiency and effectiveness of healthcare personnel, ultimately leading to improved patient care outcomes. The integration of artificial intelligence

(AI) has become increasingly prevalent in various domains of healthcare, including health services management, clinical decision-making, predictive medicine, patient data analysis, and diagnostics. Notwithstanding its remarkable accomplishments, artificial intelligence continues to be perceived as an enigmatic entity, commonly referred to as a black box.

Explainable AI (XAI) in the domain of healthcare encompasses a suite of technologies that empower medical professionals with the capacity to comprehend the underlying rationales behind the decision-making processes of AI systems. This elucidation of the "why" behind the system's choices serves as a crucial component in fostering trust, transparency, and interpretability within the healthcare AI landscape[5].

Put simply, this elucidates the fundamental reasoning of the artificial intelligence system. Numerous state-of-the-art black-box models often employ suboptimal or perplexing variables in order to attain their impressive performance levels. An illustrative example involves the utilisation of a deep learning network that has been meticulously trained on a cohort of individuals afflicted with asthma, with the invaluable guidance and expertise of medical professionals. Regrettably, this particular network erroneously deduced a significantly lower mortality rate associated with pneumonia cases. The utilisation of extraneous data, such as the positional attributes of the scanner, was observed in another instance of a deep learning architecture employed for pneumonia detection in x-ray images. In a subsequent iteration, a novel framework was devised to discern patients with varying degrees of risk by leveraging x-ray data, wherein hardware-related features were harnessed to forecast the likelihood of adverse outcomes[6]. These illustrative instances demonstrate the insufficiency of relying solely on the dependability of the models. The current landscape of artificial intelligence research necessitates the development of supplementary frameworks that cultivate a sense of assurance, notably in the form of explainable AI. With the progressive proliferation of explainable artificial intelligence (AI) within the realm of healthcare, it is reasonable to anticipate a heightened level of convergence between this burgeoning technology and wearable devices. Wearable devices have emerged as a valuable reservoir of patient data, as they diligently track and record individuals' health metrics throughout the course of their daily activities. Explainable AI has the capability to harness.

2.1 Advantages in Healthcare

- ***Improved Accuracy:*** Through the elucidation of the underlying mechanisms employed by an AI system to derive a specific inference, explainable artificial intelligence (XAI) methodologies foster a culture of transparency within the realm of healthcare procedures. The augmentation of openness and comprehension, consequently, engenders elevated levels of trust. Furthermore, the elucidations provided by eXplainable Artificial Intelligence (XAI) methodologies can facilitate the monitoring of the influence exerted by various components on the predictive capabilities of an AI system.
- ***Model Improvement:*** In the realm of AI research, a crucial aspect lies in the continuous enhancement of models. These models, which are imbued with the ability to learn from vast amounts of data, possess the remarkable capability to generate accurate forecasts. At times, it is observed that the prevailing norms and regulations may prove to be fallible, leading to the generation of imprecise prognostications. The utilisation of eXplainable Artificial Intelligence (XAI) methodologies facilitates the assessment of acquired rules, thereby enabling the identification of errors and subsequent enhancement of models.

- ***Enhanced precision:*** Through elucidating the underlying processes employed by an AI system, explainable artificial intelligence (XAI) in the realm of healthcare has the potential to augment the precision of medical diagnoses and therapeutic protocols. This facilitates the comprehension of physicians and other medical practitioners regarding the rationale behind the recommendation of a particular diagnostic strategy, thereby fostering enhanced precision and efficacy in the subsequent therapeutic interventions.
- ***Knowledge transfer:*** Explainable Artificial Intelligence (XAI) serves as a conduit for the seamless transfer of knowledge from AI systems to medical personnel, thereby empowering them to leverage this knowledge in order to optimise patient outcomes. As an illustration, an AI system has the capacity to unearth a heretofore unacknowledged correlation between a given symptom and a specific ailment. Through the elucidation of its underlying reasoning, the AI system engenders a paradigm shift within practitioners, illuminating novel interconnections that can be harnessed for subsequent diagnostic endeavours.[7]
- ***Early Detection:*** The utilisation of Explainable Artificial Intelligence (XAI) holds great potential in the realm of early detection, as it possesses the capability to meticulously scrutinise vast quantities of patient data. By doing so, XAI can uncover intricate patterns that may elude the discernment of medical professionals, thereby facilitating timely diagnoses of various illnesses. Explainable Artificial Intelligence (XAI) possesses the remarkable capability to elucidate the underlying factors contributing to an individual's susceptibility to specific diseases. This advanced system achieves this feat by employing a rigorous process of analysis and inference, enabling it to discern the intricate patterns and correlations within vast amounts of data. Consequently, XAI is able to provide comprehensive explanations regarding its conclusions, shedding light on the precise reasoning behind its determinations. Moreover, XAI goes beyond mere diagnosis, as it is equipped to propose proactive measures aimed at averting the onset of these diseases. By leveraging its extensive knowledge base, XAI offers valuable insights into preventive strategies, empowering individuals to take informed actions in safeguarding their health.

3 XAI Methods

The efficacy of explainable artificial intelligence (XAI) techniques is contingent upon several key factors. These include the specific model employed, the domain within which the problem resides, the desired degree of interpretability, and the intended recipients of the explanations. Various explainable artificial intelligence (XAI) methodologies exhibit distinct trade-offs concerning their accuracy, fidelity, and comprehensibility.

Explainable Artificial Intelligence (XAI) is an encompassing term denoting a collection of methodologies and approaches that are designed with the noble objective of furnishing lucid and comprehensible explanations for the discerning human mind regarding the intricate decision-making processes executed by artificial intelligence (AI) and machine learning (ML) models. These elucidations serve to facilitate comprehension for users and stakeholders, shedding light on the underlying rationale behind model predictions, thereby fostering trust and promoting transparency in the realm of artificial intelligence systems. Numerous eXplainable Artificial Intelligence (XAI) methodologies have been devised with the aim of attaining this objective.

3.1 Commonly used XAI methods

There are several different approaches that XAI takes, some of them are as follows:

- Local Interpretable Model-agnostic Explanations (LIME):

LIME is a well-known method that is used to produce local explanations for black-box machine learning models. Using perturbed data points, it creates a simple model that is easily interpretable (for example, a linear model) in order to get an approximation of the decision boundary that surrounds a particular instance.

The LIME approach consists of a number of critical steps, which are as follows:

- Data Manipulation: The first step that LIME does in the process of generating local explanations is to pick the instance that requires an explanation. After that, the data for this instance is altered in order to generate a collection of cases that are comparable to one another but have some subtle differences.
- Model Prediction: The black-box model is applied to each instance that has been disrupted, and the associated predictions are derived from this process. The black-box model is regarded like a "black box" in the sense that we do not need to know its internal architecture or parameters. This allows us to approach the model as if it were a "black box."
- Local Data Weighting: In LIME, weights are assigned to the perturbed instances based on how closely they resemble the initial instance. Higher weights are assigned to instances that occupy a location in the feature space that is geographically closer to the initial instance. This indicates that these instances are of greater significance when it comes to approximating the local behaviour of the black-box model.
- Interpretability Model: After that, a more straightforward interpretable model is educated with the help of the perturbed examples and the black-box model predictions corresponding to those instances. The type of interpretable model that is utilised may be anything as straightforward as a decision tree or a linear model. The goal of the interpretable model is to provide a close approximation of the behaviour predicted by the black-box model in the immediate vicinity of the instance of interest.
- Local Explanation: Once the interpretable model has been trained, it can be used to create explanations for the prediction of the initial instance. This may be done once the model has been trained. The coefficients or decision rules of the interpretable model are utilised to bring attention to the features and the various contributions that each feature had to the forecast that the black-box model had made for that particular instance.

- **SHapley Additive exPlanations (SHAP)**

The SHAP (SHapley Additive exPlanations) framework is rooted in the principles of cooperative game theory, a branch of mathematics that deals with the analysis of strategic interactions among multiple agents. By leveraging this theoretical foundation, SHAP is able to provide valuable insights into the inner workings of machine learning models. One of the key outputs of the SHAP framework is the assignment of a SHAP value to each individual feature. These values serve as indicators of the relative importance or contribution of a particular feature towards a given prediction. By quantifying the impact of each feature, SHAP enables researchers and practitioners to gain a deeper understanding of the factors driving model predictions. The proposed framework offers a comprehensive and cohesive structure for elucidating the outcomes generated by machine learning models across various domains.

- **Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) Plots:**

The utilisation of Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots has become increasingly prevalent in the field of artificial intelligence research. These visualisations serve the purpose of providing a comprehensive understanding of the average impact that one or a select few features have on the predictions generated by a given model. By employing these plots, researchers are able to gain valuable insights into the intricate relationships between features an Partial Dependence Plots (PDPs) provide a comprehensive overview of the average impact of a single feature on the model's predictions. On the other hand, Individual Conditional Expectation (ICE) plots offer a more granular perspective by illustrating the individual effect of a feature on each specific instance.

- **Global Surrogate Models**

Global surrogate models refer to interpretable models that have been trained with the purpose of approximating the predictions made by a black-box model. Surrogate models possess the advantageous characteristic of enhanced interpretability, thereby facilitating the acquisition of valuable insights pertaining to the behavioural patterns exhibited by black-box models.

- **Decision Trees and Rule-Based Models**

Decision trees and rule-based models possess an inherent characteristic of interpretability. Decision-making processes are effectively represented by artificial intelligence researchers as a sequence of uncomplicated rules or nodes within a tree structure, thereby facilitating comprehension and interpretation.

- **Attention Mechanisms**

Attention mechanisms are widely employed in deep learning architectures, with a particular emphasis on their application in the realm of natural language processing. The salient aspects of the input that the model prioritised during its cognitive deliberation are duly emphasised.

Methods for assessing feature importance, such as permutation feature importance and feature importance derived from tree-based models (e.g., Gini importance), facilitate the ranking of features according to their respective contributions to the predictive capabilities of a given model.

- **Feature Importance Methods**

Methods for assessing feature importance, such as permutation feature importance and feature importance derived from tree-based models (e.g., Gini importance), facilitate the ranking of features according to their respective contributions to the predictive capabilities of a given model.

- **Rule-based Explanations**

Rule-based explanations offer explicit and interpretable rules that articulate the reasoning behind the predictions made by the model. The aforementioned rules are derived from the internal representation of the model.

- **Contrastive Explanations**

Contrastive explanations, in the realm of artificial intelligence research, involve the process of comparing a model's prediction for a given instance with the prediction it would have generated had certain features possessed alternative values. This technique allows for a deeper understanding of the model's decision-making process and sheds light on the impact of different feature values on the final prediction. This facilitates the comprehension of users regarding the impact of varying feature values on the output of the model.

- **Anchors**

Anchors, in the context of machine learning, can be conceptualised as concise and interpretable if-then statements that serve as adequate conditions for the predictions made by the model. The authors offer succinct elucidations pertaining to the precise outputs of the model.

4 Interpretable Machine Learning Model

Interpretable Machine Learning (ML) models, colloquially referred to as Explainable AI (XAI) models, have been meticulously crafted to offer comprehensible and transparent elucidations for their prognostications. These models play a pivotal role in situations where the transparency and interpretability of the model are of utmost importance, as observed in domains like medical diagnosis, finance, and legal systems. Presented below are a few illustrations of interpretable machine learning (ML) models that have been employed in the realm of explainable artificial intelligence (XAI).

1. ***Decision trees***, a prevalent model in the field of artificial intelligence, are highly regarded for their interpretability. These models possess the ability to dissect complex datasets by systematically partitioning the data into a series of hierarchical decisions. This hierarchical decision-making process is elegantly represented as a tree structure, which further enhances the comprehensibility of the model. In the context of the tree structure, it is important to note that every individual node is inherently associated with a specific feature. Furthermore, it is crucial to acknowledge that each branch emanating from said node signifies a decision that is contingent upon the aforementioned feature.
2. ***Random Forests***, a prominent ensemble technique, are composed of numerous decision trees. The utilisation of this particular approach yields enhanced predictive capabilities while simultaneously preserving a degree of interpretability. By scrutinising the constituent trees, one can discern the significance of features and the rationale behind decision-making processes.
3. ***Generalised Linear Models*** (GLMs) represent a distinguished category of linear models that effectively broaden the scope of simple linear regression [15] by incorporating diverse target variables, such as binary or count data. The interpretability of linear models is frequently observed owing to their inherent linearity.
4. ***Linear regression***, a fundamental statistical modelling technique, is characterised by its simplicity and interpretability. This model effectively establishes a linear association between the input features and the target variable.
5. ***Logistic regression***, akin to its linear regression counterpart, is a statistical modelling technique employed for the purpose of binary classification tasks. Its interpretability is derived from its inherent linearity, which allows for a clear understanding of the relationship between the input variables and the predicted outcome [16].

6. ***Rule-based models***, encompassing rule-based decision systems and expert systems, leverage human-readable rules as a means to generate predictions and exhibit a commendable level of interpretability [17-18].
7. ***Additive models***, a class of predictive models, leverage the combination of various elementary functions, such as linear, spline, or tree-based functions, to generate accurate predictions. By integrating these simple functions, additive models are able to capture complex relationships and patterns within the data, ultimately enhancing their predictive capabilities. Interpretability can be significantly enhanced, particularly in the context of linear components.
8. ***Partial Least Squares (PLS)*** is a regression technique that has gained considerable attention in the field of artificial intelligence research. It offers a valuable solution for addressing the challenges posed by multicollinear data, a common issue encountered in various domains. PLS has demonstrated its efficacy in producing results that are not only statistically sound but also readily interpretable, making it a preferred choice among researchers and practitioners alike.
9. ***Naive Bayes***, a probabilistic model rooted in Bayes' theorem, has emerged as a prominent tool in the realm of text classification. Its popularity stems from its ability to effectively handle a wide array of classification tasks pertaining to textual data. The inherent simplicity of the subject matter facilitates its interpretability.
10. ***The LASSO***, also known as the Least Absolute Shrinkage and Selection Operator, is a linear regression technique that combines the benefits of variable selection and regularisation. By incorporating both of these aspects, the LASSO method offers a comprehensive approach to modelling and analysis. L1 regularisation, also known as Lasso regularisation, possesses the remarkable property of inducing sparsity in the coefficient space. This property endows the model with enhanced interpretability, as it allows certain coefficients to be precisely zero.
11. ***Elastic Net model*** is a sophisticated linear regression technique that effectively integrates both L1 and L2 regularisation methods. By striking a delicate balance between variable selection and regularisation, the Elastic Net model offers enhanced interpretability compared to conventional linear regression models.
12. ***Symbolic AI***, also known as symbolic reasoning systems, is a paradigm in artificial intelligence that encompasses the representation of knowledge through the use of symbols and rules. Human-readable explanations of the decision-making process are permitted by these entities [18-19].

The interpretability of these models exhibits a spectrum of degrees, and the selection of a model is contingent upon the particular utilisation scenario and the desired level of transparency. Achieving a harmonious equilibrium between interpretability and performance is of utmost importance when employing interpretable models within the realm of Explainable AI applications. It is imperative to meticulously deliberate upon the inherent trade-offs that arise in such scenarios.

4.1 Comparative Study of Four Models

Naive Bayes, a probabilistic classification algorithm rooted in Bayes' theorem, is renowned for its simplicity and effectiveness. The term "naive" is employed in this context to denote the underlying assumption that the features within the dataset exhibit conditional independence with respect to the class label. Despite the initial assumption of naivety, it is noteworthy that

Naive Bayes exhibits remarkable performance in numerous practical classification endeavours, particularly in the domains of text classification and spam filtering. The Least Absolute Shrinkage and Selection Operator (LASSO) is a regression technique that introduces a penalty term to the conventional ordinary least squares (OLS) regression. The incorporation of an L1 penalty into the regression coefficients serves a dual purpose, facilitating both feature selection and regularisation. The L1 penalty, also known as the Lasso regularisation, exerts a powerful influence on the coefficients of a model by driving some of them to precisely zero. This remarkable property endows the L1 penalty with the ability to perform feature selection, thereby simplifying the model.

The Elastic Net algorithm represents a significant advancement in the field of linear regression, building upon the foundations laid by the LASSO technique. By integrating both L1 (LASSO) and L2 (Ridge regression) penalties, Elastic Net achieves a harmonious balance between feature selection and regularisation, resulting in enhanced predictive performance and improved model interpretability. The integration of L1 and L2 penalties in a regularisation framework offers a harmonious equilibrium between the virtues of feature selection, as exemplified by the LASSO method, and the enhanced treatment of correlated features, akin to the principles underlying Ridge regression. The utilisation of Elastic Net proves to be advantageous in scenarios where the dataset exhibits a substantial quantity of features, and a subset of these features demonstrates a notable degree of correlation.

Symbolic AI, colloquially referred to as classical AI or rule-based AI, is an esteemed discipline within the realm of artificial intelligence that operates on the foundation of explicit rules and symbols. This paradigmatic approach entails the representation of knowledge and the subsequent decision-making process through the utilisation of meticulously crafted rules and symbols. This particular methodology entails the representation of knowledge through the utilisation of logical rules, thereby facilitating the process of reasoning through the application of deduction and inference rules. This stands in stark contrast to the machine learning methodologies that acquire knowledge of patterns and interdependencies through the analysis of data. Symbolic AI, once a dominant paradigm in the early stages of AI development, has witnessed a notable shift in recent times. This shift has been propelled by the emergence of machine learning, deep learning, and neural networks, which possess the remarkable capacity to acquire knowledge autonomously from vast amounts of data, thereby obviating the necessity for explicit rule-based systems.

Naive Bayes, a renowned probabilistic classification algorithm, has garnered significant attention in the field of artificial intelligence research. It leverages the power of probability theory to classify data points with remarkable accuracy. On the other hand, LASSO and Elastic Net, two prominent regularisation techniques, have emerged as indispensable tools in the realm of linear regression. These techniques effectively combat overfitting by imposing constraints on the model's coefficients.

5. Applications of XAI

Machine learning and artificial intelligence (AI) possess immense potential, both in the present and the future, to revolutionise nearly every facet of the medical field. Nevertheless, the issue of opacity in AI applications has become progressively troublesome in various domains, extending beyond the realm of medicine. The aforementioned phenomenon is particularly conspicuous in situations where users are required to decipher and comprehend the generated outcomes of artificial intelligence systems[8]. The concept of Explainable AI (XAI) encompasses the provision of a coherent and comprehensible rationale, enabling users to gain insight into the underlying factors that have contributed to a system's generation of a specific output. The resulting output possesses the potential for interpretation within a predetermined context. In recent years, there has been a growing body of evidence showcasing the remarkable efficacy of Artificial Intelligence (AI) systems across various domains.[6] These systems have consistently demonstrated their ability to produce exceptionally precise outcomes[8]. Machine Learning (ML) models that exhibit commendable performance, characterised by minimal occurrences of false positives and false negatives, such as Deep Learning networks, Support Vector Machines, or ensemble methods (e.g., Random Forest), are, nevertheless, regarded as black box models. In addition to their principal output, which typically involves anomaly detection, event prediction, or failure anticipation, these systems tend to provide minimal elucidation regarding the underlying methodologies employed to attain such outcomes. Furthermore, it is worth noting that certain machine learning models exhibit a rather disconcerting phenomenon. These models, while demonstrating commendable performance during the training and test phases, are unfortunately plagued by an enigmatic bias present within the available data. Consequently, when deployed in real-world scenarios, these models falter and struggle to generalise effectively. There are two notable challenges that arise from this situation. Firstly, the reliability of the machine learning model's output is called into question. Secondly, additional inquiries are necessary to confirm, pinpoint, and comprehend the underlying issue that prompted the machine learning model's response. The field of Explainable Artificial Intelligence (XAI) has emerged as a prominent research domain, aiming to tackle the aforementioned challenges. Consequently, it has garnered significant attention across various disciplines and industries. Given the escalating need to enhance the transparency of intricate systems for the purpose of comprehensibility and safeguarding the rights of end-users, explainable artificial intelligence (XAI) possesses the capability to facilitate heightened acceptance and dependability of AI systems[9].

The domain of Clinical Decision Support Systems (CDSSs) presents a compelling case for the integration of Explainable Artificial Intelligence (XAI) techniques. These computational systems provide assistance to medical professionals in their clinical decision-making processes. However, the lack of explainability in these systems can potentially result in challenges related to either excessive or insufficient reliance on their outputs. The provision of comprehensive explanations regarding the derivation of recommendations will enable practitioners to engage in more sophisticated decision-making processes, thereby potentially leading to outcomes that are not only more nuanced but also capable of preserving human lives in critical scenarios. The imperative for explainable artificial intelligence (XAI) within clinical decision support systems (CDSS), as well as the broader medical domain, is underscored by the imperative for ethical and equitable decision-making. It is crucial to recognise that AI models, when trained on historical data, have the potential to perpetuate and reinforce historical actions and biases that warrant scrutiny and examination. In the realm of artificial intelligence research, it is observed that tabular data processing explainable Artificial Intelligence (XAI)-enabled systems have gained significant prominence. These systems, which facilitate the interpretation and understanding of the underlying mechanisms driving their decision-making processes, have become widely prevalent in the literature. Conversely, it is worth noting that XAI-enabled Clinical Decision Support Systems (CDSS) specifically designed for text analysis purposes are comparatively less prevalent in the existing body of scholarly work. Numerous studies have elucidated the advantageous implications of Explainable Artificial Intelligence (XAI)[6]. These investigations have highlighted notable merits, including the augmentation of decision confidence among clinicians. Furthermore, XAI has demonstrated its capacity to generate hypotheses pertaining to causality, thereby fostering an environment of heightened trustworthiness and acceptability for the system. Consequently, these findings have underscored the potential for XAI's seamless integration into clinical workflows. Clinical Decision Support Systems (CDSS) represent a class of computer systems that have been meticulously designed to augment the delivery of healthcare services. These systems have recently witnessed the integration of Machine Learning (ML) techniques, which are being effectively harnessed to further enhance their development and functionality.

The process industry pertains to the intricate and multifaceted domain of converting discrete and distinct raw materials into their ultimate and consummate manifestations as finished products. The proliferation of Artificial Intelligence (AI) systems within various industries has resulted in a notable enhancement of production efficiency, a commendable reduction in energy consumption, and a significant improvement in operational safety. Notwithstanding the considerable level of automation, the involvement of human agents and their cognitive faculties continues to hold significance and indispensability in facilitating the requisite operational processes[9]. The utilisation of Explainable Artificial Intelligence (XAI) within the process industry presents a notable set of challenges, primarily stemming from the multifaceted demands that emerge from a diverse range of AI end-users and AI application scenarios. The process industry, also known as the manufacturing sector, encompasses a range of industries that are involved in the transformation of raw materials into finished products. This transformation is achieved through the application of specific recipes or formulas, which dictate the precise steps and parameters required to convert the input materials into the desired output. The process industry distinguishes itself from other sectors by its focus on the systematic and controlled manipulation of materials, rather than the assembly of pre-existing parts. Illustrative instances encompass sectors such as oil and gas, chemical, pulp and paper, metal, cement, or food and beverage industries. The operational aspects of these processes pertain to the routine activities carried out within the production facility[8]. These activities encompass the vigilant oversight and regulation of the process itself, the diligent monitoring and upkeep of the equipment, the strategic planning and scheduling of production activities, and the ongoing pursuit of enhancing the production process[10]. This pursuit may involve implementing alterations to the recipe or refining the control processes, all with the aim of achieving continuous improvement. The operational procedures within these various industries exhibit a remarkable degree of automation, yet it remains imperative to acknowledge the indispensable contributions made by human operators, engineers, and maintenance personnel. A control room operator plays a pivotal role in ensuring the optimisation of operational processes, guaranteeing both productivity and effectiveness. Their primary objective is to uphold the highest standards of product quality, while simultaneously adhering to the operational and business requirements that govern their domain [3].

6.Limitations of XAI and Future developments and trends

While the concept of Explainable Artificial Intelligence (XAI) holds great promise in enhancing transparency and interpretability in AI systems, it is imperative to acknowledge and address the inherent limitations associated with this approach. Firstly, it is imperative to acknowledge the absence of a universally accepted norm pertaining to the desired outcomes of explainability[10]. The requirements for comprehending AI models vary between developers and users, encompassing technical comprehension, as well as norms and decision-making procedures, contingent upon the particular domain. In addition to the manifold advantages of Explainable Artificial Intelligence (XAI), it is imperative to acknowledge the existence of certain limitations. These limitations encompass concerns pertaining to data privacy and security, the intricate nature of AI models, the potential for human bias, and the challenge of ensuring user comprehension. It is imperative to acknowledge the extant constraints of explainable artificial intelligence (XAI) in the present context. Several notable constraints of Explainable Artificial Intelligence (XAI) can be identified.

The domain of explainable artificial intelligence (XAI) encompasses various approaches and methodologies that often exhibit computational complexity, necessitating substantial computational resources and processing capabilities for the generation and interpretation of the valuable insights and information they offer. The computational complexity associated with XAI poses a significant obstacle when it comes to real-time and large-scale applications, thereby constraining the utilisation and implementation of XAI within these particular domains [19-21].

The prevailing challenge in the field of explainable artificial intelligence (XAI) lies in the limited scope and domain-specificity exhibited by various approaches and methods. It is crucial to acknowledge that these techniques may not possess universal applicability or relevance across the entire spectrum of machine learning models and applications[10]. The inherent limitations in the scope and domain-specificity of Explainable Artificial Intelligence (XAI) pose a formidable challenge, impeding the widespread adoption and application of this cutting-edge technology across various domains and use cases.

One of the primary challenges in the domain of Explainable Artificial Intelligence (XAI) pertains to the absence of standardisation and interoperability[11]. Presently, various XAI approaches and methodologies employ disparate metrics, algorithms, and frameworks. This heterogeneity poses a significant hurdle in effectively comparing and evaluating these approaches, thereby impeding the widespread adoption and deployment of XAI across diverse domains and applications.

The inherent constraints associated with explainable artificial intelligence (XAI) pose formidable obstacles that impede the widespread adoption and implementation of this cutting-edge technology across various domains and applications. The deployment of Explainable Artificial Intelligence (XAI) presents a multitude of advantages, as well as certain obstacles and constraints. However, it is imperative to recognise that the significance of Explainable Artificial Intelligence (XAI) cannot be disregarded, as it signifies a pivotal advancement towards attaining enhanced and accountable implementations of AI models in forthcoming endeavours.

As the Future developments and trends the trajectory of Explainable AI (XAI) appears to be highly auspicious, propelled by an escalating need for transparency and confidence in the realm of AI systems. With the increasing ubiquity of Artificial Intelligence across diverse domains, it is imperative to acknowledge the pivotal significance of Explainable Artificial Intelligence (XAI) in guaranteeing the responsible and ethical deployment of AI systems. Future advancements in the field of artificial intelligence are expected to primarily revolve around the refinement and development of deep learning models that possess increased interpretability[11]. This pursuit aims to enable a deeper understanding of the inner workings of these models, thereby facilitating their comprehension and trustworthiness. Additionally, there will be a strong emphasis on enhancing the interaction between humans and computers in the context of explainable artificial intelligence. The implementation of standardised evaluation metrics will play a pivotal role in facilitating the process of benchmarking and effectively comparing explainable artificial intelligence (XAI) techniques. Furthermore, the seamless incorporation of eXplainable Artificial Intelligence (XAI) within the broader framework of AI pipelines, coupled with the progressive advancement of tailor-made interpretability tools, will undoubtedly expedite the pragmatic deployment of XAI in diverse real-world scenarios[12]. As the field of Explainable Artificial Intelligence (XAI) progresses, it is poised to play a pivotal role in facilitating the democratisation of AI. By enabling users to comprehend and make well-informed choices grounded in the predictions generated by Artificial Intelligence systems, XAI holds the potential to cultivate a more transparent and reliable AI ecosystem[1].

Increased adoption of XAI methods for the burgeoning prevalence of Explainable AI (XAI) techniques can be attributed to the growing need for enhanced transparency, trustworthiness, and accountability within the realm of artificial intelligence systems. The necessity for explanations in AI decision-making is propelled by a confluence of factors, namely regulatory requirements, user trust, and fairness concerns. Explainable Artificial Intelligence (XAI) plays a pivotal role in facilitating and enhancing critical applications by ensuring transparency and interpretability. By adhering to legal obligations, XAI fosters a regulatory-compliant environment, thereby mitigating potential risks and liabilities[10]. Moreover, XAI contributes to optimising business outcomes by providing actionable insights and facilitating informed decision-making processes. The utilisation of human-in-the-loop artificial intelligence (AI) systems is facilitated, thereby enhancing the overall performance and capabilities of the AI models. This approach capitalises on the advancements in research and education, thereby fostering continuous improvement in the field of AI. Furthermore, the integration of enhanced explainable AI (XAI) tools further amplifies the benefits derived from this paradigm, enabling a more transparent and interpretable AI ecosystem. The practical applications of this technology in real-world scenarios serve as compelling evidence of its inherent value, while the prevailing public sentiment places a significant emphasis on the ethical implications that accompany its deployment. In the broader context, the burgeoning recognition of responsible AI underscores the significance of interpretability in constructing dependable and advantageous Artificial Intelligence (AI) systems, thereby propelling the ascendance of Explainable AI (XAI) across various sectors and use cases. There are several compelling rationales for endowing explainable artificial intelligence (XAI) systems with the capacity for justification.

Foster user confidence in the system's capabilities is the primary motivation for pursuing Explainable AI (XAI) lies in the user's desire to establish a sense of trust and confidence in the AI system. This particular factor is consistently cited as the most prominent rationale behind the pursuit of XAI.

Safeguarding against bias for ensuring the mitigation of bias is a paramount concern in the realm of artificial intelligence research. Bias, whether explicit or implicit, has the potential to significantly impact the fairness and equity of AI systems. Consequently, it is imperative for researchers. The imperative of transparency arises from the need to scrutinize the utilization of data by an AI system, particularly in relation to its potential to engender inequitable consequences.

Meeting regulatory standards or policy requirements ensuring compliance with regulatory standards and policy requirements is a critical aspect of any AI research endeavor. Adhering to these guidelines is essential to promote ethical and responsible AI development. By meeting regulatory standards, researchers can demonstrate their commitment to transparency, fairness, and accountability in their work. Improving system design for the consideration of transparency and explainability holds significant weight when it comes to the implementation of legal frameworks pertaining to artificial intelligence systems. The empirical evidence substantiates the assertion that a given product or service has successfully satisfied the requisite regulatory benchmarks[10].

Estimating risk, robustness, and vulnerability to increase the capacity of the AI system to provide explanations can greatly facilitate the comprehension of its decision-making processes by developers, thereby enabling them to discern the underlying rationale behind its behavior and subsequently devise appropriate enhancements. The task at hand involves the estimation of risk, robustness, and vulnerability. This multifaceted endeavour necessitates a comprehensive understanding of the potential hazards, the ability to withstand and recover from adverse events, and the susceptibility to harm or damage. By employing rigorous methodologies and advanced analytical techniques, we aim to quantify

Understanding and verifying the outputs from a system for gaining a comprehensive understanding of the underlying mechanisms governing a system's operation can prove instrumental in effectively assessing and quantifying associated risk factors.[8] The imperative nature of conducting thorough evaluations arises when a system is introduced into an unfamiliar environment, thereby engendering uncertainty regarding its efficacy. The incorporation of this particular attribute holds significant merit within the realm of safety-critical domains, such as airports, automated teller machines (ATMs), and aircraft operations.

The task at hand involves comprehending and validating the outcomes generated by a given system. The incorporation of interpretability techniques within the evaluation process of an AI system can prove to be highly advantageous in the context of verifying its outputs[8]. This elucidates the intricate interplay between modelling decisions and the selection of data, both of which exert discernible influence on the resultant outcomes of the system.

There is a growing recognition that the ethical implications of black box artificial intelligence systems render the notion of the ends justifying the means untenable. Historically, the mere acquisition of insights sufficed as evidence for the feasibility of an AI platform. However, in the realm of critical use cases, it has become imperative for AI platforms to demonstrate their capacity to elucidate the underlying rationale behind their decision-making processes. New methods and approaches are highly probable that forthcoming epochs will witness the emergence of novel methodologies and approaches, which shall bestow upon us the ability to procure machine learning models that are endowed with enhanced transparency and interpretability. The methodologies and approaches under consideration may be predicated upon diverse principles and perspectives, thereby affording a more holistic and nuanced comprehension of the inner workings of machine learning models.

Increased demand and adoption are plausible to anticipate a surge in demand and widespread adoption of explainable AI in the forthcoming era. This projection is rooted in the growing realisation among various entities and individuals regarding the inherent merits and favourable outcomes associated with the utilisation of transparent and interpretable machine learning models[4]. The surge in demand and subsequent adoption of artificial intelligence (AI) has the potential to catalyse the advancement and implementation of novel explainable AI techniques and methodologies. Consequently, this could pave the way for the proliferation of AI applications that possess a greater degree of transparency and comprehensibility, thereby amplifying their reach and influence across various domains [22-24].

Regulatory and ethical considerations for the forthcoming era will undoubtedly witness an intensified emphasis on the regulatory and ethical dimensions of explainable artificial intelligence (AI), as a growing number of entities and individuals acknowledge the profound implications and ramifications that this cutting-edge technology may entail[13]. The exploration of this avenue holds the potential to engender the emergence of novel benchmarks, protocols, and conceptual frameworks pertaining to the domain of explainable artificial intelligence[14]. Moreover, it stands poised to furnish a structural scaffold that facilitates the conscientious and principled deployment of this technology, thereby ensuring its responsible and ethical utilisation [25,26].

In light of these forthcoming advancements and emerging patterns in the realm of explainable artificial intelligence, it is highly probable that they will yield substantial ramifications and find diverse utility across various domains and contexts. The aforementioned advancements possess the potential to engender novel prospects and complexities within the realm of explainable artificial intelligence (AI), thereby exerting a profound influence on the trajectory of this cutting-edge technology.

7 Conclusion

Explainable AI, although still in its nascent phase, holds significant promise in fostering consumer confidence and trust in the deployment of increasingly sophisticated AI applications. In the interim, enterprises shall be compelled to address the lacunae and engage in conjecture regarding the algorithms employed by the artificial intelligence system to arrive at a decision. The realm of explainable artificial intelligence (AI) is teeming with a multitude of forthcoming advancements and emerging trends. These developments possess the potential to yield profound consequences and find practical utility across diverse domains and applications. Several noteworthy advancements and emerging patterns in the realm of explainable artificial intelligence (AI) warrant attention.

References

1. L. Merrick and A. Taly, *The Explanation Game: Explaining Machine Learning Models Using Shapley Values*, vol. 12279 LNCS. 2020.
2. D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G. Z. Yang, “XAI-Explainable artificial intelligence,” *Sci. Robot.*, vol. 4, no. 37, pp. 4–6, 2019, doi: 10.1126/scirobotics.aay7120.
3. R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, “A historical perspective of explainable Artificial Intelligence,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 11, no. 1, pp. 1–21, 2021, doi: 10.1002/widm.1391.
4. G. Vilone and L. Longo, “Notions of explainability and evaluation approaches for explainable artificial intelligence,” *Inf. Fusion*, vol. 76, no. April, pp. 89–106, 2021, doi: 10.1016/j.inffus.2021.05.009.
5. F. K. Dosilovic, M. Brcic, and N. Hlupic, “Explainable artificial intelligence: A

survey," *2018 41st Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2018 - Proc.*, pp. 210–215, 2018, doi: 10.23919/MIPRO.2018.8400040.

6. Y. Zhang, Y. Weng, and J. Lund, "Applications of Explainable Artificial Intelligence in Diagnosis and Surgery," *Diagnostics*, vol. 12, no. 2, 2022, doi: 10.3390/diagnostics12020237.
7. I. Tiddi and S. Schlobach, "Knowledge graphs as tools for explainable machine learning: A survey," *Artif. Intell.*, vol. 302, p. 103627, 2022, doi: 10.1016/j.artint.2021.103627.
8. A. M. Antoniadi *et al.*, "Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review," *Appl. Sci.*, vol. 11, no. 11, pp. 1–23, 2021, doi: 10.3390/app11115088.
9. A. Kotriwala, B. Kloepper, M. Dix, G. Gopalakrishnan, D. Ziobro, and A. Potschka, "XAI for operations in the process industry - Applications, theses, and research directions," *CEUR Workshop Proc.*, vol. 2846, 2021.
10. R. Machlev *et al.*, "Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities," *Energy AI*, vol. 9, no. May, p. 100169, 2022, doi: 10.1016/j.egyai.2022.100169.
11. A. Saranya and R. Subhashini, "A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends," *Decis. Anal. J.*, vol. 7, no. February, p. 100230, 2023, doi: 10.1016/j.dajour.2023.100230.
12. B. R. Rajagopal and C. T. Solutions, "Analysis of Current Trends , Advances and Challenges of Machine Learning (Ml) and Knowledge Extraction : From Ml To," vol. 58, no. September, pp. 54–62, 2022.
13. J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, "Evaluating XAI: A comparison of rule-based and example-based explanations," *Artif. Intell.*, vol. 291, p. 103404, 2021, doi: 10.1016/j.artint.2020.103404.
14. J. M. Mendel and P. P. Bonissone, "Critical Thinking about Explainable AI (XAI) for Rule-Based Fuzzy Systems," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 12, pp. 3579–3593, 2021, doi: 10.1109/TFUZZ.2021.3079503.
15. Dr. Shikha Verma, Mrs. Priya Mishra, Mr. Prashant Verma, Mrs. Meghna Gupta. (2022). STUDENT's BEHAVIOR AND REGULARITY EFFECTS RESULT. Journal of Pharmaceutical Negative Results, 6995–7003.
16. Kumar, K.; Pradeepa, M.; Mahdal, M.; Verma, S.; RajaRao, M.V.L.N.; Ramesh, J.V.N. A Deep Learning Approach for Kidney Disease Recognition and Prediction through Image Processing. Appl. Sci. 2023, 13, 3621.
17. Gade, K., Geyik, S. C., Kenthapadi, K., Mithal, V., & Taly, A. (2019, July). Explainable AI in industry. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 3203-3204).
18. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.
19. Emmert-Streib, F., Yli-Harja, O., & Dehmer, M. (2020). Explainable artificial intelligence and machine learning: A reality rooted perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(6), e1368.
20. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science robotics*, *4*(37), eaay7120.
21. Avasthi, S., Chauhan, R., & Acharjya, D. P. (2023). Extracting information and inferences from a large text corpus. *International Journal of Information Technology*, *15*(1), 435-445.
22. Avasthi, S., Prakash, A., Sanwal, T., Tyagi, M., & Yadav, S. (2023, January). Tourist reviews summarization and sentiment analysis based on aspects. In *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 452-456). IEEE.
23. Rachha, A., & Seyam, M. (2023). Explainable AI In Education: Current Trends, Challenges,

And Opportunities. *SoutheastCon 2023*, 232-239.

24. Hasib, K. M., Rahman, F., Hasnat, R., & Alam, M. G. R. (2022, January). A machine learning and explainable ai approach for predicting secondary school student performance. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 0399-0405). IEEE.
25. Maurya, Rajesh Kumar, and Sanjay Kumar Yadav. "Ensemble Classification Approach for Cancer Prognosis and Prediction." In *International Conference on Biologically Inspired Techniques in Many-Criteria Decision Making*, pp. 120-135. Springer, Cham, 2019.
26. Maurya, R. K., Yadav, S. K., & Tewari, P.(2020). Use of Artificial Intelligence (AI): A Developing Assessment Techniques for Study of Tumor Diversity from Gene Expression. Psychology and Education. Volume-57, Issue-9,pp.1781-1785 ,ISSN: 0033-3077.