A Precise review of XAI and a roadmap of XAI for education

Karuna Wangkhem[1]

[1]NIELIT IMPHAL,
NIELIT Imphal, Akampat
karuna.wangkhem18@gmail.com

Abstract*: *From Artificial Intelligence as a subject belonging to Computer Science, the advancement has reached to a point where AI is a disciplinary branch currently. From simple soft computing algorithms to machine learning algorithms, smaller artificial neural network to deep learning, AI has been explored extensively. At this point, the black-box nature of the technology is started to be seen as a little crevice left behind against reaching its utmost potential. So, the concept of explainable AI arises where XAI gives a route or an explanation that can be decrypted in human understandable form. Since some few years there is a big rise in the research towards XAI as it has a promising potential towards providing transparent, accountable, unbiased, authentic and ethically aligned result. In this paper, the taxonomies, current state and trends, applications areas, opportunities and limitations of XAI are discussed from papers from 2020. A special roadmap of XAI in the education field is drawn out in the later section of the paper.

***Keywords*: Artificial Intelligence, Deep Learning, Machine Learning, Explainable AI, Education.**

(Article history: Received: 1June,2024 and accepted 22 Nov,2024)

1. INTRODUCTION

Millennials have actually gone through different phases of computer evolutionary generations. From computer era to information era then comes the internet phase then the infamous Artificial Intelligence generation arrived in the human civilization. AI is such a revolution that the fifth-generation computer era is simply called as AI generation. Even this AI also, it has gone through multiple form of transitions. In a holistic manner, AI can be defined as the programming of Computer system to have human like intelligence in the form of cognitive and aptitude. Therefore, from a whole discipline of computer science of hard computing there came a shift towards soft computing that can handle more than just a single input-output system. From there Machine Learning becomes a subset of AI. Machine learning is more or less, technique to identify patterns and try to give predictions with the data trained. It gives the prediction through its own heuristic learning process where the basic classification of ML is supervised learning and unsupervised learning. Now everything was fine with ML but when it comes to different types of data like structured and unstructured data, ML has its shortcomings. Therefore, under its subset again, a different Learning called Deep Learning arises where all types of data can be handled. DL basically come from the concept of artificial neural network where the model tries to imitate how human brain works which is in a mesh structure. DL is expressed generously through Convolutional neural network and other gradients. Here again, a little problem comes with DL due to its lack of capability to provide the meaning to its predictions. Even if the predictions are very accurate, the interpretability of the system is questionable. Basically, the system is fully based on black-box model where the whole process and the path to the result is completely blind for the user. As the demand for its reliability and accountability are at stake, the new subset of Explainable Artificial Intelligence or XAI was born. XAI is a subset of the very superset of AI where the predictions are given some form of explanation of why the prediction is so and how did the prediction reach to the prediction with the given inputs. In simple language, a white-box model of the whatever AI model implemented is the core concept of XAI. In this paper, a section is dedicated for the various application areas of XAI after reviewing different relevant papers. A section is again added for all the taxonomies regarding XAI. The discussion section will pinpoint out a particular area of concern from all the XAI reviews and proposed a simple manifesto. The main significance of the proposed roadmap can be highlighted in concern with a very big issue which is currently facing by todays' generation. Education- a field where human cognitive abilities are supposed to be enhanced, is taking a very different route. Speaking from the current scenario, education instead of tickling the human mind, is becoming more or less AI generated questions and answers. The problem is not the ability of AI to give answer rather the issue is AI generating answers or results without any

reference value and transparency of how the result is so. The mentioned section will try to delve a little bit and break the ice.

2. XAI AND ITS APPLICATION AREAS

There is a huge demand of explanation in AI Models in many fields in the current era of Artificial Intelligence. Looking from that perspective, this paper tried to segregate some application domains of XAI and some credit is given to some areas where the scope of XAI can be expanded. The exploration of XAI in the respective fields vary greatly though. Some fields like healthcare, finance and agriculture are pushing extensively towards XAI. Following is the description of some of fields where XAI can be really a boon.

A. Healthcare

Saranya et. al. [11] did an extensive survey on the status of XAI in the field of healthcare. In this paper, the dominance is proven through a statistical report showing 43% of XAI research are done in the healthcare field and the remaining 57% is shared by all the miscellaneous fields. C.C. Yang et.al [12] have actually used clinical trials data and applied Deep Learning and experimented to embed XAI to give an explanation from the result. Here, V. Jahmunah et. al. [13] used Gradient weighted Class Activation Maps to visualize the prediction provided for myocardial infarction disease from the Deep Learning models of DenseNet and Convolutional Neural Network or simply CNN. Here, the Class Activation Maps (CAM) provided a unified average interpretation or explanation of the predictions provided by both DenseNet and CNN. Additionally, instead of using DL, Dave et.al [37] used Machine Learning Techniques and used post-hoc XAI model-agnostic techniques on the dataset of heart disease. The dataset is taken from UCI ML Repository and the machine learning algorithm used is XGBoost ML technique heart disease dataset and explainability was implemented using Feature-Based Technique. One of the points that can be taken is both LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) is implemented to illustrate the local interpretability and global explainability using XAI. In addition to feature-based techniques, example-based techniques were shown through Anchors, Counterfactuals, Counterfactuals guided by prototype, Kernel SHAP on the heart disease dataset. And the short overview of integrated gradient was also shown using Fashion MNIST dataset. Shoes and shirt dataset were used to apply the positive and negative attributions of integrated gradients. Overall, in healthcare section, from medicine, psychology, lung disease, heart disease, brain disease to teleophthalmology, it can be found that XAI is being explored in a good pace of research journey. A major breakthrough was found when the statistics showed that AI models are more accurate in diagnostic capability however the big loophole was the transparency. Less transparency comes with the price of ethical issues. Therefore, XAI comes into the picture when transparency was the need of the hour.

B. Finance

Precisely, finance is a sector which governs the momentum of human development. Therefore, it is to be noted that AI made the path for the level of development finance sector has achieved right now. However, a big trade-off comes with the AI as the amount of fraud and scams increases with the automation of monetary market. Again, AI plays the role of moderators of all the shenanigans. It can help in predicting credit card fraud, investment scams and many more. One of the very popular one which is shown in the top of search result is loan through AI and using kernel SHAP XAI technique. For example, the details of loan, its interest, acceptance or rejection criteria and its fulfilment will be explained to the loanee. One of the noteworthy works done on this field is prediction of credit card default application. Tanusree et.al [40] combined clustering of hidden network and the TREPAN decision tree. The dataset used was from UCI repository and python was used and a level of interpretability was added to the neural network model. As in finance sector, wrong prediction can be catastrophic, a level of necessary robustness and stability has to be provided and it can be through explanation of the prediction and its route to the prediction.

C. Environmental Science (especially agriculture)

Right now, throughout the world, a special curriculum framework is increasingly added to educate people of the sustainable development and the call for the hour to protect our earth from all the degradation and pollution. Therefore, it is again the high time to inculcate the potentials of AI in this field. For climate management, forest carbon stock, forest inventory models are some areas where AI has already reached however further room for research is still there in this field. For agriculture, leaf disease detection, pest control and critical prediction of pesticide are some examples. GoogleNet, ResNet is found to be the extensive source for agricultural AI models especially for Deep Learning models.

D. Cyber Security (especially social media)

Accept it or not, it is a matter of fact that every person living in this world right now are cyber citizen in one or another form. Digital footprints are there for every single individual even an infant who just got delivered. Social media somehow became a part and parcel of today's lifestyle. News is spread more rapid than a wildfire, in this global village, a news can reach to the opposite part of the planet in the matter of seconds not even minutes. Here, the catch is the reliability of the news being generated in millions per second. Fake news is like a day today term nowadays. There are many fake news detection systems using various AI models however, news is something which carry a sentiment value to itself. The news which is circulated first always carry a higher priority to retain. Therefore, in order to actually dissipate a fake news, an explanation of why and how a news is authentic or fake is much necessary step. And yes, BERT (bidirectional-encoder-representations from transformers) is seen to be used often in fake news detection with LIME. Similarly, H.Mehta et.al.[24] used HateXplain dataset to detect hate speeches. They used BERT and ANN combined to get around 93.5% accuracy. LIME was added for local interpretability. Similar works regarding scams like deepfake and all can also be seen to be under research.

E. Education

There is a huge framework specifically designed for education called AIED (Artificial Intelligence for Education) to enrich the educational quality and all the stakeholders involve in the educational transfer process. The framework of AIED is progressing rapidly and AIED tools can totally act as intelligent teaching assistants providing a new learning experience like never before. Some of the popular Large Language Model (LLMs) like GPT or commonly known as ChatGPT can literally generate enough verifiable educational content nowadays. To put it briefly, it is marketed in such a way that education without AIED in today's era especially post COVID is shown to be frown upon or outcasted. However, the key point is AIED is somehow losing the human touch or the essence due to the black-box nature of the LLMs to be precise. Here, a huge abyss is already created due to excessive automation and training that it is sometimes seen to be out of control. One example that can be noted is many users have encountered situations where some LLMs give out results with citations but the citations were found to be non-existent. Somewhere down the line, how the results are obtained needs to be little more transparent than the way how it is currently. Fiok et.al [15] pinpointed out how XAI can be used in education and training. Here, two types are extracted how human skills can be hone through AI. Type 1 is competition between human and AI. One example provided was competition in chess like games using AlphaZero AI model. After that Type 2 was explained as aid of AI during the educational process of the human. The example taken was how copilot kind of assistance being provided by the AI models like GPTs. Slow but steadily, education domain is also getting attention from XAI research.

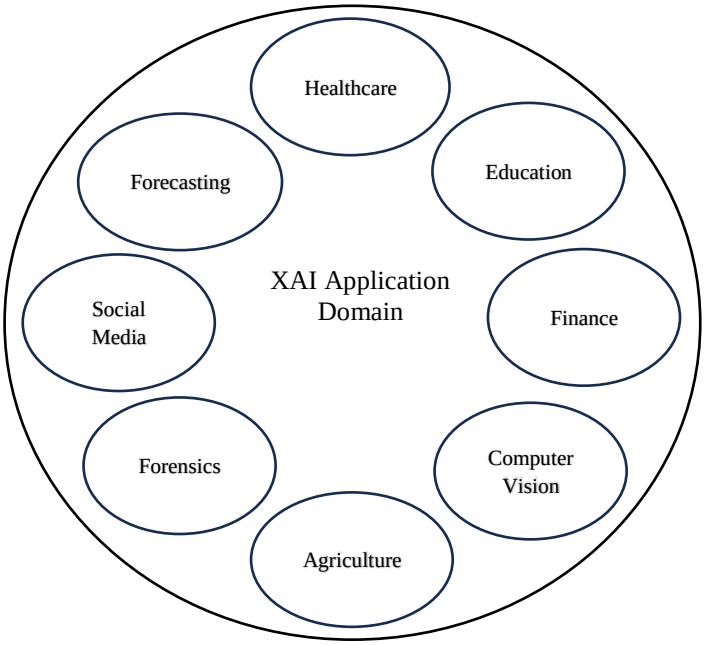

Fig. 1. XAI Application Domain

Table 1: Important XAI points from different authors

Author	Year	Important points
Tanusree et.al. [40]	2020	Provision of human interpretability to deep learning is established
Tjoa et.al. [38]	2020	Explainable AI was envisioned from the medical perspective
Merry et.al. [31]	2021	A novel mental model fordefining XAI was explored
Mehta et.al. [24]	2022	Hate speech detection from social media was implemented using XAI
Saranya et.al. [11]	2023	The application domain of XAI- opportunities and trends are thoroughly analysed
Liu et.al [4]	2024	XAI was studied from the angle of education as an application domain

3. XAI CLASSIFICATION METHODS

XAI can be classified in different types based on different methods. XAI can be seen to be classified in most papers as *ante-hoc and post-hoc*. Even ante-hoc and post-hoc are described as explainability models also. However, after summarizing all the classic taxonomies of XAI, XAI can be classified based on the:

a. Explanation technique
b. Approach
c. Scope

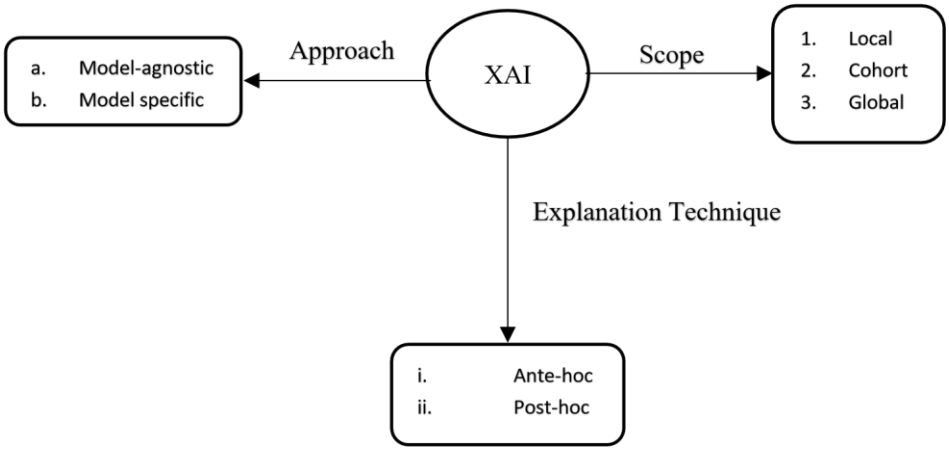

Fig. 2. XAI Classification basis

A.Ante-hoc vs post-hoc

The central theme of explainable AI is clearly the explanation provided to the user in the human understandable form. However, some Machine Learning techniques are self-explanatory in some extent. Therefore, based on its techniques, XAI explanations can be of two types- the one that is intrinsically explainable while the other kind being a technique where explanations are provided after the training.

Ante-hoc explanations are provided through the incorporation of interpretability techniques into the model or architecture or the learning process or even the intrinsic property of the algorithm itself. A well-known algorithm is Decision Tree (DT), despite its simple and easy approach, DT is still an overused algorithm for its ante-hoc explainability property. It gives a comparatively transparent approach where the result or the outcome can be traced back as an intrinsic characteristic. One of the evident examples of ante-hoc explanation can be seen in finance sector. Results and predictions are generally in the form of visualized expression or just a tabular data format consequently providing an ante-hoc explanation. It possesses this characteristic due to its implementation of Rule-based models of Machine Learning. Some more models that can be roughly counted under this category would be

-Trepan Reloaded

-LIME (Local Interpretable Model-agnostic Explanations)
-GAM (Generalized additive Models)
-BRL (Bayesian Rule Lists)

On the other hand, post-hoc explainability model is being applied after the training process. It is going to analyse and interpret the decision-making process after the prediction and going to provide all the relevant insights of how the result is given as output as so. One of the biggest advantages is it doesn't depend on the type of model used for training but rather proving a way to bring an extensive amount of AI models in the exposure of a non-black-box AI prediction or result. It clearly gives the property of transparency yet making it more flexible. Most importantly, it is to be noted that recently, post-hoc models are in the trend and being explored than ever before. Some of worth mentioning models under post-hoc would be

-SHAP (Shapley Additive Explanations)
-Anchors
-LIME (Local Interpretable Model-agnostic Explanations) -Counterfactuals etc.

Table 2: Differences between ante-hoc and post-hoc

	Ante-hoc	***Post-hoc***
Key difference	It is intrinsically explainable.	Explainability is achieved after training.
Mechanism	Interpretability is directly incorporated into the learning process allowing it to give an explanation of the result provided.	It interprets the decision-making process of a trained machine learning model to provide the route to its result.
Applicability	It is suitable for small and structured dataset where relationships are there.	It can also fit complex and unstructured dataset
Accuracy vs Interpretability	Sometimes, accuracy can be compromised for interpretability.	Accuracy is the one resulting from the learning model whereas interpretability is provided by the add-on XAI model.
Examples	Decision Trees, Rule Based Models	LIME, SHAP

B.Model specific and model-agnostic

XAI can again be classified on the basis of approach to provide its explanation. One can be exclusively for a specific model while one can be generic relatively. From this perspective, XAI can be either *model specific or model-agnostic.* Model specific approach are techniques specifically oriented for a particular model. They are going to give explanations solely focusing on the given model. In the initial phase of XAI, model specific approach was the only way naturally. However, due its inflexible nature, the current XAI advancement is no longer supporting towards a model specific behaviour. Nevertheless, it stood a milestone pioneer stand to advancement of XAI techniques.

In the contrast, model-agnostic approach follows techniques that treats a learning model as just a black-box model and it is gives the explanation in the typical post-hoc manner. That's how it provides independence from a specific model and flexibility. Currently, model-agnostic techniques are getting enough push through and most research are more or less towards model agnostic only.

Table 3: Differences between Model specific and Model-agnostic

	Model specific	***Model-agnostic***
Key difference	Completely based on a particular model.	Explanation is not just based on a single model.
Mechanism	Explanation is provided solely based on a particular model.	Explanation are provided without accessing its internal working model and simple perceiving any model as a black-box only.
Applicability	It is applicable for a single or similar model.	It is applicable to all the generic models
Relevance vs Flexibility	It is high in accuracy and relevance but inflexible.	Accuracy is again based on the model; relevance is comparatively lesser but versatile.
Examples	Bayesian networks, multilayer perceptron approaches	LIME, SHAP

C.Scope of XAI techniques

All the different types of XAI techniques can be classified into different types on the basis of how much scope and span of dataset they can cover and explain. Taking into account of their capabilities, the scope can broadly be local, cohort and global. When local scope is discussed, the derivation of the explanation is from a single instance of a model. It focuses on understanding a particular prediction or a singular result. The individual characteristics leading the result is pin pointed out in the form of explanation or interpretation. One of the models that exemplifies local scope would be LIME.

While, cohort scope focuses on a group of similar instances of a model. It is used to give explanation of the model's behaviour for a particular subset of data instead of individual point or the entire dataset. One XAI techniques that can be considered under cohort is PDPs (Partial Dependence Plots). Likewise, the scope that covers the entire dataset would be called as global. Its aim is to give explanation for the result it is providing or otherwise the behaviour of the model or the algorithm. It is pretty similar to learning techniques where patterns and rules are observed and mined. Global surrogate models are feature based model explainability technique exclusively for global scope XAI explanations.

Table 4: Scope of XAI Techniques

	Local	***Cohort***	***Global***
Key difference	Based on a single instance of a model.	Based on similar instances of a model.	Based on the overall dataset.
Mechanism	Explanation is provided solely based on a singular instance of the model.	Explanation is provided based on group of similar instances of a model.	Explanation is provided based on behaviour of the model.
Applicability	It is applicable for individual level decision.	It is applicable for group level analysis.	It is applicable for model level audit.
Granularity	Granularity is high	Granularity is intermediate.	Granularity is low.
Examples	LIME	Partial dependence Models	Global surrogate models

Model-agnostic XAI techniques can be roughly classified into the following types again:

- Feature-based model explainability techniques
- Example-based Model explainability techniques
- Rule-based Model explainability techniques

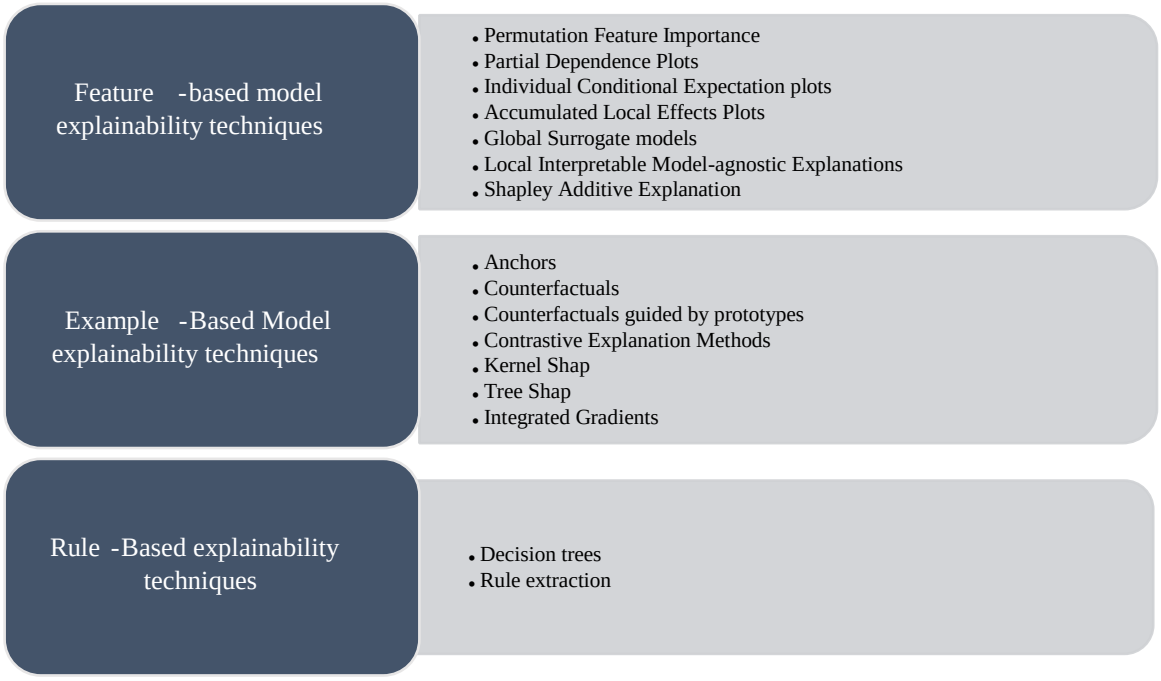

Fig. 3. Model-agnostic explainability techniques

a. Feature- based model Explainability techniques

Feature-based model explainability techniques focus on identifying and quantifying the importance of input features in model predictions.

Permutation Feature Importance is a model-agnostic technique where the specific feature is randomized and break the relationship of the feature and the target then, will quantify the level of degradation in the performance. If there is very less degradation, it can conclude and explain that the feature was of obviously less importance. Let us view this with the following code/ pseudo-code:

```
X= np.random.permutation(X)
y=model.predict(X)
accuracy=accuracy_score(Y,y)
importance=base_accuracy-accuracy
feature_importances_value=importance
```

Partial Dependence Plots can work with both white-box in XAI and the typical black-box learning technique. It is generally used to plot the relationship between input features and the target when some features are hold constant and some varying, so the name partial dependence. Example can be seen from the given code/pseudo-code:

```
pdp_feature=pdp.pdp_isolate(model=model,dataset=X,model_features=feat
ure_names,feature='selected_fearure' )
```

Individual Conditional Expectations (ICE) plots are PDPs in local scope format. It is going to give the visualize result of the relationship between the specific feature and the prediction. The idea is extracted from PDP only that can be clearly seen in the following code example:

```
result=partial_dependence(model,X,features=[selected_feature],kind='t
ype',grid_resolution)
```

Accumulated Local Effects plot is again an explainability technique that divides the features into range and then an interval and then explains how the feature influences the result within the interval or the local scope. ALE plots can be illustrated as below with the usage of libraries of explainers:

```
from alibi.explainers import ALE ale =
ALE(model.predict, feature_names)
ale_explanation = ale.explain(X)
```

Global surrogate models are one of the ante-hoc model-agnostic technique where the learning model are exploited with its interpretability property. Therefore, here decision trees, linear regression or similar kind of algorithms are implemented. The capability of interpretation of the Decision Tree can be shown with the following code/pseudo=code:

```
y_predict = model.predict(X)
global_surrogate_model=DecisionTreeClassifier(depth,random_state)
global_surrogate_model.fit(X,                                    y_predict)
y_predict_surrogate=global_surrogate_model.predict(X)
fidelity= accuracy_score(y_predict,y_predict_surrogate)
```

Local Interpretable Model-agnostic Explanations (LIME) is a model-agnostic technique which is completely for local scope. It tries to estimate the impact of a singular data sample locally through fitting into the model being applied. It takes an interpretable model like DT or linear regression and approximates with the model locally. A simple code that can describe the working of LIME is provided below:

```
exp.explain_instance (X.iloc[target].values,model.predict_proba)
```

SHAPley Additive Explanation (SHAP) follows a very popular game theory and gives a global scope of explanation. This popular game theory was coined by Shapley in 1953 where he used a concept of a game where multiple players are contributing in a game. The contributions are assumed to be similar relatively and the results are termed as "pay-outs". So, here dataset instance equals players, prediction equals to the pay-outs, task of prediction equals to the game, difference between the actual of local instance and the average of the unified instances equals to the gain. The biggest advantage of SHAP is it can give both local explanation and the global explanation. The local explanation can be visualize using the logic of the following code:

```
shap.force_plot(explainer.expected_value,shap_values[instance,:],X.iloc[in
stance,:])
```

In the similar manner the same thing can be translated into global explanation using the very SHAP with the following code snippet:

```
shap.force_plot(explainer.expected_value,shap_values[:2000,:],X.iloc[:2000
,:])
shap.summary_plot(shap_values,X)
```

b. Example- based model Explainability techniques

Example-based model explainability techniques give a qualified value of how much an instance contributes to a model's output or predictions.

Anchors have cohort scope which mean that its number of instances taken is greater than that of LIME but usually lesser than what is generally called to be global. Anchors are basically the range of conditions which are narrowed down for prediction and explanation. They provide rules for a specific prediction and explain why the prediction is so. The rules are if-then in general case. The biggest advantage of anchors is its boundaries are clear and adapts perfectly to the training model. However, the disadvantage comes again when the instance is not yet trained and rules might not be easy to interpret. A small code snippet for anchor usage is given below:

```
anchor=explainer.explain(X[instance])
print(anchor.anchor)
```

Counterfactuals are the explanations that provides the alterations occurred with the least amount of change happening to the feature values. For counterfactuals, the ideal is binary datasets. The rule applied here is what-if condition. For example, what will happen to Y if X is altered (even a minute change). A rough code for generating counterfactuals is given below:

```
cf=dice.generate_counterfactuals(query_instance,total_CFs,desired_class)
print(cf.cf_examples_list.final_cfs_df)
```

Counterfactuals guided by prototypes are sophisticated counterfactuals in the sense that examples which are representative in nature i.e. protypes are provided to increase the interpretability of the model used. The explanations become more realistic and model become a little more transparent. A rough layout of the program of counterfactuals guided with prototype is mentioned below:

```
prototypes= kmeans.cluster_centers instance=np.array([])
prototype=prototypes[np.argmin(np.linalg.norm(prototypes-instance,
axis=1))]
new_query=instance.copy()
new_query= prototype[0] + i
```

Contrastive Explanation Model (CEM) is again a model-agnostic XAI model that gives explanations in such a way that the reason of not being the result that is not predicted is discussed. That's why it is said to give contrastive explanations. There are two types of CEM explanations- Pertinent Positives (PP) and Pertinent Negatives (PN). PP tries to find the features that are compulsory to reach to the prediction that is given out whereas PN tries to find the features that are necessarily missing to reach to the output or the prediction.

```
from alibi.explainers import CEM explainer
test=CEM(model.predict_proba,              mode)
explainer.fit(X,d_type,shape)
explanation=explainer.explain(test,**paramaters)
print(explanation.PN)
print(explanation.PP)
```

Kernel SHAP is a variation of SHAP which can again give local and global interpretation. In this case, it doesn't train the model with all the plausible permutations rather the missing values in the training are obtained from a generated formula. The code snippet can be modified from the traditional SHAP as follows:

```
explainer=shap.KernelExplainer(model.predict,shap.kmeans(X))
values=explainer.shap_values(X[:])
shap.summary_plot(shap_values, X[:],feature_names)
```

Tree SHAP provides exact SHAP values which is optimized for tree-based models. The exception of tree SHAP is it is generally model specific XAI technique. And du to its model specific nature, it has high computational efficiency. The little modification is there in the kernel SHAP code that is shown below:

```
explainer= shap.TreeExplainer(model.predict,shap.kmeans(X))
values = explainer.shap_values(X[:])
shap.summary_plot(shap_values, X[:],feature_names)
```

Integrated Gradients which are also known as Path Integrated Gradients or Axiomatic Attribution again belong to a model-agnostic XAI technique that is different from the above techniques as it deals with Deep Learning models or otherwise complex machine learning models. As it covers Deep Learning, it can handle both structured and unstructured datasets. It can have two types of attributions – positive attributions and negative attributions. Positive attributions are the attributions that helps to reach to the decision or the prediction. On the other hand, negative attributions are the attributions that tried to negate the decision that is eventually reached.

```
ig_attribution=integrated_gradients(model,input_image,base)
```

c.Rule- based model Explainability techniques

Rule-based model explainability techniques use logical rules and derivatives to describe model behavior.

Decision Trees are originally Machine Learning algorithm where the internal nodes are the features, branches are the outcomes and the leaf nodes are the predictions or the outputs. It has the property of interpretability because it follows the if-then-else rule. The biggest problem faced is its vulnerability towards overfitting.

```
Rbm1=DecisionTreeClassifier(max_depth,random_state)
Rbm1.fit(X,Y)
rules=export_text(Rbm1,feature_names)
```

Rule extraction is an umbrella term used to cover all the rule-based model explainability models that has the property of intrinsic interpretability due to its approximation to its prediction using rules. One of the most generic rules applied is if-then rule. The biggest problem is it doesn't provide AI transparency to that extent that would be counted as appreciable.

Table 5: Comparison of XAI Techniques

XAI Technique	**Local Explainability**	**Global Explainability**
Permutation Feature Importance	×	✓
Partial Dependence Plots	✓	✓
Individual Conditional Expectation plots	✓	✓
Accumulated Local Effects Plots	✓	✓
Global Surrogate models	×	✓
LIME	✓	×
SHAP	✓	✓
Anchors	✓	×
Counterfactuals	✓	×
Counterfactuals guided prototype	✓	×
Contrastive Explanation Methods	✓	×
Kernel SHAP	✓	✓
Tree SHAP	✓	✓
Integrated Gradients	✓	×

4. DISCUSSION

In previous sections, it is thoroughly discussed about the application domain and the taxonomies of XAI. Through all this, this paper will focus on an area where there is a huge potential of XAI exploration- Education. Khosravi H et.al [16] talks about a beautiful framework where XAI meets education called XAI-ED framework. Here, all the stakeholders to be involved are identified and the benefits towards them are also analysed. Pitfalls are also pointed out along with their avoidance clue. Fiok et.al. [15] briefly talked about basic goals of XAI in education quoting reliability, transparency, privacy, causality, usability, fairness and trust being the main points.

This paper also would try to put forth a strategic roadmap or a manifesto for XAI in education. The illustration can be shown as follows:

A. Source

The source of the very idea of XAI in education can be sorted out like the following:

- The basic aim was to enrich the educational process through computer-aided training where human performance can be enhanced.
- The second source comes from AIED where the users are ignorant of how the decisions and the results are coming.
- The third source is ironically due to the developers being unaware of the actual theme of the subject.

B. Stakeholders

The main stakeholders involved in this context are as follows:

- *Learners* are simply the student candidates who are receivers of the educational process.
- *Parents* are the guides of the wards as the maximum wards are minors in the educational process.
- *Teachers* are the senders of the educational contents or data generators
- *Technologists* are the tech people who are going to be directly involved in the implementation of the XAI.
- *Educational researchers* are the people who is going to research and input their work towards educational technology.
- *Policy makers* are those authority who have the power to exercise legal and political curriculum framework.

C. Goals or Benefits

The main benefits of XAI in education would be like:

- The very basic benefit would be eventually the increase in AI literacy.
- There would be sense of accountability due to the provision of answers with why and how explanations.
- Reliability comes as a subordinate victory as the system comes with objective-causality pair.
- One benefit can also be flexible and better interaction between student and teacher.

D. AI Models

Some popular AI models applicable are:

- Decision Trees (DT) are easy and simple models that doesn't make the system complex rather has its ante-hoc explainability property.
- Clustering methods will be useful to create clusters of similar features where explainability can be easily added.
- Generalized Additive Model (GAM) can be used to find the relationship between different types of input variables.

E. XAI Models

XAI models and their key points are already in the above sections and though all can be experimented, some models that has promising output would be like:

- LIME would surely be ideal for local explanation of a feature-based input.
- Similarly, for a global explanation SHAP would come handy and Python SHAP is in quite in application.
- Counterfactuals will work satisfactorily in situation where explanation is needed for a minimal change.

F. Approach

Different types of approach can be implemented to have XAI-ED kind of framework. Those can be summed up as follows:

- *Learner centric approach*, a technique which is the core of modern education system, can be designed in such a way that the learner is the focal point of the AIED. Learner feedback, reviews and a tentative behaviour analysis of the learner and the system would be taken into account.
- *User experience approach* would put the ease and comfort of the user to use the system as the top priority.
- *Theory-driven approach* would actually have the usage of XAI model that has the highest accuracy and completeness. The model choice would be taken into account before anything else.

- *Participatory-centric approach* is a type of approach where the involvement of maximum stakeholders would be encouraged.
- *Human Computer Interaction approach* is a design approach that would improvise how the computer system is interacting with the human. For example, for visual learner, explanation with visual representation would surely be a better HCI whereas for an audio learner, a speech assistant would be more appropriate.

G. Risk and its mitigation

Probably, everything comes with a risk inevitably. However, the fatality rate can be decrease with a good mitigation plan. Similarly for this case also, some risks are identified and mitigation plan is mentioned as follows:

- With the involvement of AI models as well as XAI models, the unified model can be complex in nature. However, it can be handled smoothly with a choice of less complex models to begin with.
- The explanations can be inaccurate as well as incomplete. Therefore, AI models with proper training should be applied for system or better accuracy value while suitable XAI techniques are to be implemented in order to avoid incomplete explanations.

5. CONCLUSION

Overall, in this paper, the evolution of XAI bypassing different levels and layers and even seasons of AI is discussed to give an overview of history of XAI. The current state and trends of XAI, from its singularity to becoming a generic model-agnostic techniques of XAI are also illustrated. Even though XAI was always there in its implicit form somehow, the direct extraction of the desired explanation that gives the transparency and trustworthiness of AI through the post-hoc techniques of XAI is also mentioned elaborately. The application areas of XAI and areas where potentials of XAI are yet to be researched and explored more, are also pin pointed out and a special focus was casted on the education field. All sorts of classifications of XAI applying all the possible approaches and methods are thoroughly discussed in the paper. A small strategic plan was surely mentioned for XAI-ED framework to be realized in a smoother way. Education was the epicentre of the discussion where a simple manifesto is discussed from the source to even maintenance. Ultimately, the great potential in the journey of research and development regarding XAI and its application areas is enlightened to an extent that hopefully my lead further exploration in the mentioned area. The joy of knowing how the answer is arrived additional to just getting the answer is the motivation of Explainable Artificial Intelligence.

REFERENCES

[1] Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I. H., Muller, M., & Riedl, M. O. (2024, May). The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-32).

[2] Atakishiyev, S., Salameh, M., Yao, H., & Goebel, R. (2024). Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access*I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] Hamida, S. U., Chowdhury, M. J. M., Chakraborty, N. R., Biswas, K., & Sami, S. K. (2024). Exploring the Landscape of Explainable Artificial Intelligence (XAI): A Systematic Review of Techniques and Applications. *Big Data and Cognitive Computing*, *8*(11), 149.

[4] Liu, Q., Pinto, J. D., & Paquette, L. (2024). Applications of explainable ai (xai) in education. In *Trust and Inclusion in AI-Mediated Education: Where Human Learning Meets Learning Machines* (pp. 93-109). Cham: Springer Nature Switzerland.

[5] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., ... & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, *106*, 102301.

[6] Rane, N. L., & Paramesha, M. (2024). Explainable Artificial Intelligence (XAI) as a foundation for trustworthy artificial intelligence. *Trustworthy Artificial Intelligence in Industry and Society*, 1-27.

[7] Rane, J., Mallick, S. K., Kaya, O., & Rane, N. L. (2024). Enhancing black-box models: advances in explainable artificial intelligence for ethical decision-making. *Future Research Opportunities for Artificial Intelligence in Industry 4.0 and*, *5*, 2.

[8] Pinto, J. D., Paquette, L., Swamy, V., Käser, T., Liu, Q., &Cohausz, L. (2024). Human-Centric eXplainable AI in Education (HEXED) Workshop.

[9] Wang, Y., Zhang, T., Guo, X., & Shen, Z. (2024). Gradient based Feature Attribution in Explainable AI: A Technical Review. *arXiv preprint arXiv:2403.10415*.

[10] Weber, R. O., Johs, A. J., Goel, P., & Silva, J. M. (2024). XAI is in trouble. *AI Magazine*, *45*(3), 300-316.

[11] Saranya, A., & Subhashini, R. (2023). A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, *7*, 100230.

[12] C.C. Yang, Explainable artificial intelligence for predictive modeling in healthcare, J. Healthc. Inform. Res. 6 (2022) 228–239, http://dx.doi.org/10.1007/ s41666-022-00114-1.

[13] V. Jahmunah, E.Y.K. Ng, Ru-San Tan, Shu Lih Oh, U Rajendra Acharya, Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals, Comput. Biol. Med. 146 (2022) http: //dx.doi.org/10.1016/j.compbiomed.2022.105550.

[14] Anand, A., Kadian, T., Shetty, M. K., & Gupta, A. (2022). Explainable AI decision model for ECG data of cardiac disorders. *Biomedical Signal Processing and Control*, *75*, 103584.

[15] Fiok, K., Farahani, F. V., Karwowski, W., & Ahram, T. (2022). Explainable artificial intelligence for education and training. *The Journal of DefenseModeling and Simulation*, *19*(2), 133-144.

[16] Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... &Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074.

[17] K. Wei, B. Chen, J. Zhang, S. Fan, K. Wu, G. Liu, D. Chen, Explainable deep learning study for leaf disease classification, Agronomy 12 (2022) 1035, http: //dx.doi.org/10.3390/agronomy12051035.

[18] T. Speith, A review of taxonomies of explainable artificial intelligence (XAI) methods, in: C. Isbell, S. Lazar, A. Oh, A. Xiang (Eds.), Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2239–2250, http: //dx.doi.org/10.1145/3531146.3534639.

[19] Gulsum Alicioglu, Bo Sun, A survey of visual analytics for Explainable Artificial Intelligence methods, Comput. Graph. 102 (2022) 502–520, http://dx.doi.org/ 10.1016/j.cag.2021.09.002.

[20] Giorgio Leonardi, Stefania Montani, Manuel Striani, Explainable process trace classification: An application to stroke, J. Biomed. Inform. 126 (2022) http: //dx.doi.org/10.1016/j.jbi.2021.103981.

[21] M. Obayya, N. Nemri, M.K. Nour, M. Al Duhayyim, H. Mohsen, M. Rizwanullah, A. Sarwar Zamani, A. Motwakel, Explainable artificial intelligence enabled TeleOphthalmology for diabetic retinopathy grading and classification, Appl. Sci. 12 (2022) 8749, http://dx.doi.org/10.3390/app12178749

[22] R.A. Zeineldin, M.E. Karar, Z. Elshaer, et al., Explainability of deep neural networks for MRI analysis of brain tumors, Int. J. CARS 17 (2022) 1673–1683, http://dx.doi.org/10.1007/s11548-022-02619-x.

[23] Minh, H.X. Wang, Y.F. Li, et al., Explainable artificial intelligence: a comprehensive review, Artif. Intell. Rev. 55 (2022) 3503–3568, http://dx.doi.org/10. 1007/s10462-021-10088-y.

[24] H. Mehta, K. Passi, Social media hate speech detection using explainable artificial intelligence (XAI), Algorithms 15 (2022) 291, http://dx.doi.org/10. 3390/a15080291

[25] R. Confalonieri, L. Coba, B. Wagner, T.R. Besold, A historical perspective of explainable Artificial Intelligence, WIREs Data Min. Knowl. Discov. 11 (1) (2021) e1391, http://dx.doi.org/10.1002/widm.1391, URL https://wires. onlinelibrary.wiley.com/doi/abs/10.1002/widm.1391.

[26] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research, Artificial Intelligence 296 (2021) 103473.

[27] M. Langer, K. Baum, K. Hartmann, S. Hessel, T. Speith, J. Wahl, Explainability auditing for intelligent systems: A rationale for multidisciplinary perspectives, in: T. Yue, M. Mirakhorli (Eds.), 29th IEEE International Requirements Engineering Conference Workshops, in: REW 2021, IEEE, Piscataway, NJ, USA, 2021, pp. 164–168, http://dx.doi.org/10.1109/REW53955.2021.00030.

[28] R. Confalonieri, L. Coba, B. Wagner, T.R. Besold, A historical perspective of explainable Artificial Intelligence, WIREs Data Min. Knowl. Discov. 11 (1) (2021) e1391, http://dx.doi.org/10.1002/widm.1391, URL https://wires. onlinelibrary.wiley.com/doi/abs/10.1002/widm.1391.

[29] A.K.M. Nor, S.R. Pedapati, M. Muhammad, V. Leiva, Overview of explainable artificial intelligence for prognostic and health management of industrial assets based on preferred reporting items for systematic reviews and meta-analyses, Sensors 21 (2021) 8020, http://dx.doi.org/10.3390/s21238.

[30] G. Joshi, R. Walambe, K. Kotecha, A review on explainability in multimodal deep neural nets, IEEE Access 9 (2021) 59800–59821, http://dx.doi.org/10. 1109/ACCESS.2021.3070212.

[31] M. Merry, P. Riddle, J. Warren, A mental models approach for defining explainable artificial intelligence, BMC Med. Inform. Decis. Mak. 21 (2021) 344, http://dx.doi.org/10.1186/s12911-021-01703-7.

[32] S.A. Fang, N.C. Tan, W.Y. Tan, et al., Patient similarity analytics for explainable clinical risk prediction, BMC Med. Inform. Decis. Mak. 21 (2021) 207, http://dx.doi.org/10.1186/s12911-021-01566-y.

[33] Z.U. Ahmed, K. Sun, M. Shelly, et al., Explainable artificial intelligence (XAI) for exploring spatial variability of lung and bronchus cancer (LBC) mortality rates in the contiguous USA, Sci. Rep. 11 (2021) 24090, http://dx.doi.org/10.1038/ s41598-021-03198-8.

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

[35] Garc´ıa, M.V., Aznarte, J.L.: Shapley additive explanations for no2 forecasting. Ecological Informatics 56, 101039 (2020)

[36] Thampi, A.: Interpretable AI, Building explainable machine learning systems. Manning Publications, USA (2020)

[37] Dave, D., Naik, H., Singhal, S., & Patel, P. (2020). Explainable ai meets healthcare: A study on heart disease dataset. *arXiv preprint arXiv:2011.03195*.

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

[39] .-Y. Chen, C.-H. Lee, Vibration signals analysis by explainable artificial intelligence (XAI) approach: Application on bearing faults diagnosis, IEEE Access 8 (2020) 134246–134256, http://dx.doi.org/10.1109/ACCESS.2020.3006491.

[40] Tanusree De, Prasenjit Giri, AhmeduveshMevawala, Ramyasri Nemani, Arati deo explainable AI: A hybrid approach to generate human-interpretable explanation for deep learning prediction, Procedia Comput. Sci. 168 (2020) 40–48.

[41] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science robotics*, *4*(37), eaay7120.

[42] Došilović, F. K., Brčić, M., & Hlupić, N. (2018, May). Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210-0215). IEEE.