From Black Box to Glass Box: A Review into Making Artificial Intelligence Explainable

D. Das, G. Lundy, S. Nikolaou, N. Patel
MSc. Machine Learning and Deep Learning, Electronic & Electrical Engineering, University of Strathclyde, Glasgow, Scotland, UK
EE986: Assignment and Professional Studies

ABSTRACT This literature review provides a comprehensive analysis of the current state of explainable artificial intelligence (XAI) tools, with a focus on recent advancements in model-specific and model-agnostic methods. The review highlights the importance of XAI in various applications and considers ethical considerations, such as the types of bias that can be introduced into a system and how XAI tools can be used to mitigate these biases. The review evaluates state-of-the-art XAI methods, evaluation and tools, including LIME, SHAP, DeepLIFT, and CAM, and presents frameworks for their use in applications like healthcare for use in improving medical diagnosis and decision-making processes, and manufacturing to optimize production processes by providing operators with real-time feedback and insights into process variables. Additionally, the review examines the business implementations associated with the adoption of XAI tools, such as the current scope of implementation and the need for companies to invest in transparent and interpretable decision-making processes. Alongside this, a simple experiment of LIME implementation has been provided to demonstrate how explainability can be embedded in a prediction provided by a neural network. Overall, the paper concludes by discussing the future challenges and opportunities in XAI research, emphasizing the need for continued development of more transparent and interpretable models, as well as effective human-machine interfaces to enhance XAI adoption.

INDEX TERMS Explainable Artificial Intelligence (XAI), Interpretability, Machine Learning (ML), XAI State-of-Art, LIME, XAI Evaluations, XAI Application, XAI Review, Explainability in business, Business Users and Decision Makers.

I. INTRODUCTION

As Artificial Intelligence has developed and advanced, humanity is having to tackle a brand-new challenge, namely the retracing and comprehension of the decisions and outputs of our algorithms and models. Many of the high-level systems currently in development or operating in real-world sectors are very complex and opaque, known as ‘black box’ algorithms. Whilst these algorithms have been very much accepted and integrated into our daily lives in places like movie recommendations on Netflix, there are still risks when blindly following and trusting the outputs of the algorithms when it is not clear how the machine has come to the decision that it has. The transparency of the decision-making process is not important for a lot of AI applications, such as the previously mentioned Netflix recommender, but it carries a lot of weight and influence when evaluating and risk assessing the use of AI models to complete such tasks as aiding to diagnose Cancer or navigate heavy machinery through a pedestrian-filled city centre or down the motorway at 70 miles per hour.

This paper will delve into the currently developing contract to these systems, Explainable Artificial Intelligence (XAI), focusing on the State-of-the-Art in the field and outlining what these methods are, how things are developing, and where they can be used.

II. BACKGROUND

Since the research into artificial intelligence began, it has been argued and debated by members of the community that any developed intelligent system should be able to explain its results.

It has however been in the past decade that AI has been brought into the spotlight with popular new research in the field of Machine Learning, having been revived by deep learning. Often performing above and beyond a human level, models that use deep learning are comprised of millions of parameters, and without the toolbox of methods that have been developed with the aim to help users and researchers understand the processes, they will very often form complex black box systems. These are very high functioning systems that may deliver amazing results but are very hard to trust within applications that will influence the way people live, and maybe even if they live at all. By highlighting pathways that were taken by the model, XAI tools can provide explanations for these decisions as well as ensuring that protected features are not negatively influencing the outcome. [1]

The field focuses on the research and development of safe, responsible AI systems for use in all sectors by implementing tools that deal with transparency and traceability to remove the black box from developing technologies and giving machines the capability to create explanations that are optimally structured for human decision-makers.

To make things simple, it can be assumed that there are two main forms that the explanation can be given. The first is an explanation that is suitable for the end user of the system that may not have any knowledge of computer science or programming, with the second being the opposite, a more comprehensive, detailed explanation that is suitable for the programmers and researchers that are developing within the toolbox. Both explanations are extremely important, yet very different and it is imperative that the system can understand the difference between the two and deliver the appropriate response when asked.

As with every artificial intelligence technique, no one XAI tool or method will be the right fit for every scenario. Just as it is important for there to be an explanation from the system, it is also important to be getting the correct form of explanation. There have already been massively important advances in the field, including tools like heatmap explanations of deep neural network classifiers, with applications in fields such as digital transformation which incorporates automation tools and robotics, sustainable living, etc.

A. HOW EXPLAINABILITY WORKS

The explainability of a machine learning model is typically the inverse of its prediction accuracy when no explanatory mechanisms are used. Meaning, the higher the prediction accuracy, the lower the model explainability. [2] This is discussed in a paper named 'Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges' by Feiyu Xu et al. The Explainable AI program at DARPA put together a chart that illustrates this effect using different types of AI models.

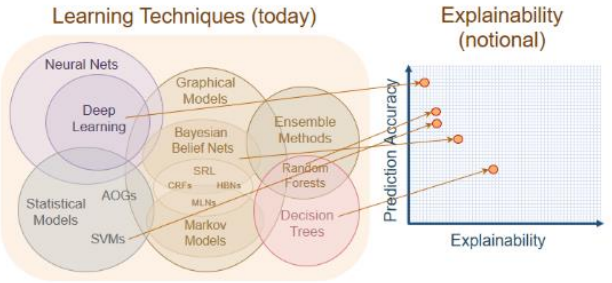

FIGURE 1. **Explainability of a model vs the precision accuracy [2]**

The illustration (figure 1) above shows that models such as decision trees have a very good level of explainability. The downside is that they have the lowest prediction accuracy of the above-shown techniques. This is because they have a graphical structure that helps to facilitate the visualisation of the decision-making process. On the other end of the spectrum, when relying on no XAI tools, the output of a DNN cannot be explained by either the developer of the system or the neural network itself, not helped by the overwhelming complexity of the learning algorithm of these models.

It is not just the learning algorithm that is required to provide an intuitive explanation, but also each individual component in the model. Parameters of a linear model could represent how strong the association between certain features and the output is. 'Explainable Software Analytics' by Hoa Khanh Dam and Truyen Tran explains a popular method of analysing a deep network through the visualisation of the neuron weights and activations. This can be through examples such as feature maps in CNNs and reoccurring activation patterns in RNNs [3]. Both techniques can also be used to estimate feature importance, which can be used to see which features from the data used to train a model are being considered as the most important and therefore are being prioritised when making decisions. Knowing which features are taking priority in a model is important because of a few reasons, including mitigating bias in an AI model, highlighting where further research may need to take place, and understanding which parts of the data should be kept or removed from the pool to potentially increase the accuracy and reduce training costs

as fewer data will need to be processed by the learning algorithm.

B. ETHICAL ARTIFICIAL INTELLIGENCE

Ethics is a huge subject within artificial intelligence and for good reason. In 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', Arrieta et al, write about the growing demand for ethical AI as well as the reluctance of humans to adopt technologies that are not interpretable and trustworthy. This topic is highly debated not only in the field of engineering and computer science but across the general population worldwide as more and more sectors are expanding to include such systems in the day-to-day operation of the world, whether that is mobile phone assistants, or software that can detect and display tumors that would be usually missed by the human eye. [4]

There are many ethical considerations that must be taken into account when designing and developing a model that could affect real lives, such as integrated bias in the learning process that carries onwards across the whole lifespan of the model. Machine learning models are a lot like humans in the way that we both learn from the information that is given to us, whether it is in a dataset in the case of AI, or taught to us by friends, teachers, parents, etc. However, unlike humans, most AI models cannot learn for themselves and the world around them, therefore making them very vulnerable to bias, accidental or otherwise.

S Richardson writes about different types of bias in a paper named 'Exposing the many biases in machine learning' [5]. In this text, 12 different biases are described after being split into the three stages of data science for the purpose of creating prediction models: capture, curation, and analysis. Figure 2 (below) shows these three stages alongside the 12 sorted biases.

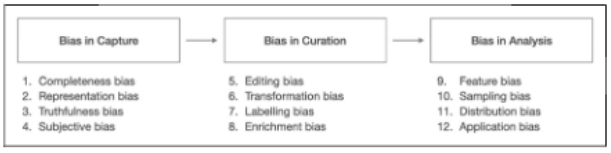

FIGURE 2. Types of AI bias in each stage of data science for creation of prediction models [5].

Explainable AI can help mitigate bias in data capture by providing insights into the data collection process and suggesting ways to collect more diverse and representative data. An example of this is applying XAI tools in an artificially intelligent hiring system for a workplace. L. Hofeditz, S. Clausen, et al, developed such a system and recorded their findings in a report named 'Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring'. The authors focused on an XAI approach that was not highly technical but instead provided a high-level explanation of how the AI system selects candidates and how it considers attributes that are regarded as sensitive. This is an example of the first type of explanation that was described previously, one that is suitable for the user over the developer. Using XAI in this way helps to build trust in the outputs as well as preventing the unethical use of an AI system. [6] The figure below (figure 3) is the research model that the authors proposed showing how XAI tools will affect the system.

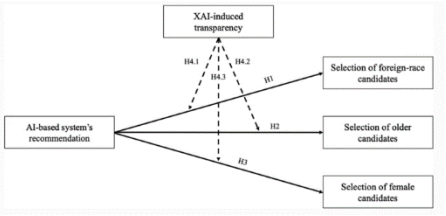

FIGURE 3. Proposed research model by L. Hofeditz et al showing how explainable tools will be used [6].

TABLE I
DATA BIASES VS FATE CHARACTERISTICS [7]

Origins	Bias	Fairness	Accountability	Transparency	Explainability
Data Source	Population data	X		X	X
	Measurement error	X			X
	Data quality chasm	X	X	X	
	Data repurposing	X	X		X
	Data augmentation	X	X	X	
Pre-Processing	Dataset shifts	X		X	
	Opaque pre-processing			X	X
	Data labeling	X	X	X	X
	Adversarial manipulations	X		X	X
	Transfer learning	X	X	X	

Within the stage of curation, XAI can help to mitigate bias by promoting transparency, accountability and interpretability as proposed in 'Establishing Data Provenance for Responsible Artificial Intelligence Systems' by K. Werder et al. The paper addresses the lack of data provenance in AI-based systems and how this can be mitigated. The table above summarises the effect that data biases have on responsible AI based on the FATE characteristics (Fairness, Accountability, Transparency, Explainability) for both capture and curation of data. The authors explain how some studies have shown to suggest that there is a potential conflict between explainability and other FATE characteristics, such as a possible trade-off between fairness and explainability [7]. Below, figure 4 shows the data provenance framework for responsible AI.

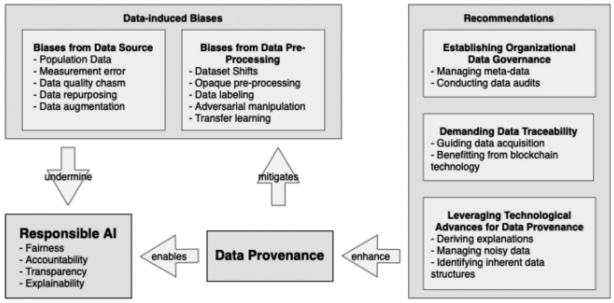

FIGURE 4. **Data provenance framework shown by K. Werder et al [7].**

XAI methods can help identify which features are biased by analyzing their contribution to the model's output. This analysis can reveal features that are highly correlated with sensitive attributes such as race, gender, or age. Once these features are identified, they can be removed or reweighted to reduce their impact on the model's predictions. They can also be used to promote fairness in machine learning models. One way to do this is through fairness constraints, which set limits on how much each feature can contribute to the model's output. By setting these constraints, it becomes easier to mitigate feature bias and ensure that the model is fair and unbiased.

In the paper 'Building Ethics into Artificial Intelligence' written by Han Yu et al, it can be seen that the principles established by the Belmont Report for behavioural sciences (1978) have been suggested as a starting point for ensuring ethical human-AI interactions. These principles are: [8]
The personal autonomy of a person should not be violated (maintain free-will when interacting with the technology).
The benefits brought about by the development of technology must outweigh the negatives.
The benefits and risks must be distributed equally across all peoples using the technology (no person should be discriminated against based on their personal background such as race, gender, and religion).
These principles serve to prove that protected characteristics should not be allowed to influence the model unless it is directly involved in the explicit purpose of the model.

III. STATE OF THE ART

A contemporary shift has been taking place from an organisation-centric to a person-centric view in varied industries, allowing the users to control the operations. Disruptive trends like big data, artificial intelligence (AI), digitalisation, human-machine interaction, analytics, computational power, and internet-of-things (IoT) have been utilised to realise and brace this shift. However, industries are currently undergoing another paradigm movement that will necessitate the application of interpretable analytics, virtual and augmented reality, smart control, and three-dimensional view models [9]. Therefore, making the operations robust, personalized, vigorous, and reason-based analytics, to derive business solutions that are highly innovative. This is possible with AI algorithms like neural networks that perform compounded operations on enormously generated data sets to gain precise predictions and detections.

Yet, the impact of AI on social and individual levels is questionable globally due to various concerns about accountability, ethics, legality, trust, transparency, and data protection involved in the use of AI especially in healthcare, in turn hindering black-box approaches, ultimately giving rise to Responsible AI [10]. To overcome these challenges, using explainable artificial intelligence (XAI) will help achieve improvement in model and result tracing since XAI enables the model's interpretability and explainability through feature extraction. XAI could be applied to varied fields like finance, logistics, human resource, and most importantly healthcare. Additionally, XAI addresses trusted analytics by its application over different decision models and allows effortless debugging and boosts the performance of trained models through an additional module which justifies the model's output decision. XAI also assists in transparent operations by defining explainability to address specific perspectives like legal, medical, technological, and financial aspects.

Moreover, XAI not only provides transparency to justify predictions of the model by AI algorithms, but it also aids in bias reduction by the system's decision overriding thereby being a fair and safe model whose decision is worth believing by people as it gives the end users exemplary explanations behind its predictions and decisions plus complies with ML/DL algorithms as per the set parameters [4]. Besides, every AI/ML system that seeks interpretability, transparency, and comprehensibility has a core objective of the explainability of a model, defined as a summary to explain the model's working, features, calculation, and final output. This explainability of models to define the summary can be of two types: Explanation by Simplification (Creates a simplified version of the model for explanation) and Explanation by knowledge extraction (Explanation through weights, biases, and I/O pairings). Explanation by simplification is a technique of XAI that can be justified as equivalent model comparison or usage of simplified models. On the contrary, the explanation by knowledge extraction can be justified as a stepwise explanation by changing the parameters with respect to the original models. Both explanations can be seen as a state of the art in most of the XAI techniques discussed further.

To begin with, this objective of explainability has led to recent development in XAI systems that can be majorly characterized by three important parts of designing an XAI model: (A) goals, (B) methodology, and (C) evaluation.

A. GOALS

The first part of designing an XAI model that is goals was justified by the Defence Advanced Research Projects Agency (DARPA). DARPA has also taken the initiative and is engulfed to design AI/ML models to address important goals that not only maintain high performance but produce human-centric explainability. These goals include trust, understandability, fairness from bias, insightfulness, causality, transferability, and privacy with data protection. In view to fulfil these goals an organization or an individual must demonstrate a better understanding of the data, problem, and methodology. Hence, to understand the data and its problem, the input and output data are categorised as numerical, pictorial, textual, and time-series [23] and for simplification the problem is subdivided into classification and regression.

B. METHODOLOGY

Noting the extensiveness of the second part of designing an XAI model which is the methodology, the following paragraphs will dive deeper into this concept to derive a better understanding of it.

The XAI model's scope has two approaches: the *local* approach - which needs the individual prediction explanation (which means articulating a section of the ML/AI System) and the *global* approach - which needs the whole model explanation (which means articulating the entire procedure of the ML/AI System) [1]. To comprehend these approaches better and help developers, as well as practitioners, identify the drawbacks of the presently available methods together with selecting the best-fit method for the model explanation, Table 1 is prepared listing the most popular XAI models. Amongst these models, Local Interpretable Model-Agnostic Explanation (LIME) and Layer-wise Relevance Propagation (LRP) were the earliest generic techniques used for explaining the decisions of complex ML models.

Correspondingly, the XAI system model design can be classified into two techniques: the *transparent* (Ante-hoc) technique - which are models that are inherently designed to explain and the *post-hoc* technique - which are models that need explanation after it is implemented. Mostly the dataset for these post-hoc approaches are Images, Tabular Data, and Time-Series and can be derived as either factual, counterfactual, or semi-factual explanations according to human understanding [18].

Under the transparent (Ante-hoc) technique certain ML/AI models can be termed as fully transparent if all three degrees of transparency are met namely, simulatability (model output can be interpreted through input data and calculation), intelligibility (all the steps of the model are interpretable), and algorithmic transparency (means an expert user understands the model by his own). These three degrees are only found in simple ML algorithms like Linear Regression, Decision Trees, K-Nearest Neighbour, and more [9]. However, presently there has been a rising interest in the algorithmic transparency model after the discovery of Fuzzy Rule-Based Systems (FRSs) which uses a knowledge base and inference engine module for model interpretability [22]. Despite both simple ML algorithms and FRSs providing transparency, these models have failed to give us high-performance and absolute results in real-time ML problems, and therefore it was necessary to use complex algorithms such as Neural Networks or Convolution Neural Networks. But these complex algorithms have absolutely no degree of transparency involved. Thus, a need for a post-hoc technique was mandatory [13].

TABLE II
OVERVIEW: TYPES OF XAI TECHNIQUES

Explainable Technique		Explanation Forms	Interpretation Types	Model Specificity	Explanation Scopes
(1) BP (2) Guided-BP (3) Deconv-Network (4) LRP	(5) deepLiFT (6) CAM (7) Grad-CAM (8) Integrated Gradients	Feature Map	Post-hoc	Specific	Local
(1) LIME (2) GraphLIME (3) SHAP (4) CLEAR (5) CERTIFAI	(6) Anchors (7) ASV (8) Shapley Flow (9) Explainable Neural-Symbolic Learning	Feature Map	Post-hoc	Agnostic	Local
(1) XGNN (2) BETA	(3) Skater	Feature Map	Post-hoc	Agnostic	Global
(1) GAM		Feature Map	Post-hoc	Specific	Global
(1) Attention		Feature Map	Intrinsic	Specific	Local
(1) CBR Solution	(2) Meaningful Perturbation	Example-Based	Post-hoc	Agnostic	Local
(1) Scoped Rules (2) DTD	(3) ProtoPNet (4) xDNN	Example-Based	Intrinsic	Specific	Local
(1) Triplet Network		Example-Based	Intrinsic	Specific	Global
(1) ICE		Textual	Post-hoc	Specific	Both
(1) VQA		Textual	Post-hoc	Agnostic	Local
(1) TVAC		Textual	Post-hoc	Agnostic	Global
(1) Image Captioning		Textual	Intrinsic	Specific	Local

The post-hoc technique from a professional perspective can be broadly divided into three types according to the application and use case as shown in Figure 5 [1]. These are Text Explanation, Feature-Based Explanation, and Examples-Based Explanation. Text Explanation of a model can be defined as the usage of a semantic description to explain the decision of the model this includes XAI techniques like Image Captioning, Testing with Concept Activation Vector (TVAC), and many more. Apart from this, the text explanation can be also characterized as a visual explanation that gives a behaviour of an overall system by a visual representation which can be helpful to

non-technical users. The Feature-Based Explanation as the name suggest presents a Saliency feature map explanation of important parameters that have the greatest effect on their output, these techniques are widely researched and provide a robust agnostic approach to any application. These include Back Propagation (BP), LIME, Generalized Class Activation Mapping (Grad-CAM), Non-linear LIME (GraphLIME), Shapley Additive Explanations (SHAP) and many more [14]. Lastly, an Example-Based Explanation usually deals with a counterfactual (what-if) approach that uses a prototype to create a similar model used for the application to provide the feature extracted during the runtime of the original model, for example, Triplet Network (TN), Prototype Part Network (ProtoPNet), and Explainable Deep Neural Network (xDNN).

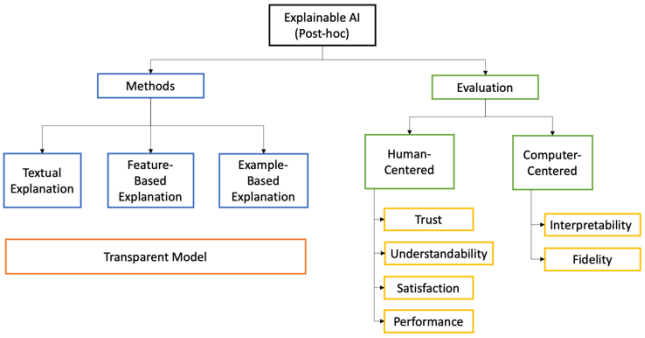

FIGURE 5. Tree Diagram for simplification of XAI post-hoc model methodology and evaluation.

When a post-hoc model explanation technique is required in an application, it is necessary to further specify the level of usage of these techniques. So, this usage level can be specified as either *model-agnostic* (no challenge of technique transferability to other ML models) or *model-specific* (applied to only a few closely related ML models). In comparison with model-specific, the use of model-agnostic techniques is more robust in different applications and makes the XAI technique more versatile to work with.

Understanding the various divisions of the XAI techniques above, now moving ahead, explained below are a few of the most cited, worked with and popular XAI models from Table-II.

1) Riberio *et al.* (2016) proposed LIME, which approximates the input and ML/DL model to provide an interpretable and trustworthy explanation of classifier output [25]. With the theory of LIME, a perturbation-based strategy called Anchors was developed by the same authors (2018), which showcases IF-THEN rules instead of surrogate models in LIME [26].

2) Another popular model was developed by Lundberg and Lee (2017) who presented SHAP, this technique assigns SHAP values to relevant features for individual predictions [27]. Similarly, Chen *et al.* (2019) proposed generalized DeepSHAP, which uses a Deep-Explainer to estimate SHAP values layer-wide in DL models [28].

3) Shrikumar *et al.* (2019) proposed Deep Learning Important FeaTures (DeepLIFT) which uses the activation function of neurons as a scoring method to evaluate the prediction in Multi-layer Perceptrons (MLPs) and DNNs [24].

4) While Class Activation Mapping (CAM) was developed by Zhou *et al.* (2015) on Convolution Neural Network (CNN) to identify the class dependencies of the most important parts of images through saliency maps to predict the decision of classification [19]. Based on this technique Selvaraju *et al.* (2020) proposed Grad-CAM which uses gradients of the target class to generate saliency maps [20].

Most of these post-hoc techniques can now be found in the academic framework with libraries such as Interpret, Alibi, Aix360, Dalex and Dice [21]. Also, open-source frameworks are now available such as SHAP lib, LIME lib, Shapash, ELI5, InterpretML, OmniXAI and many others to explore for development purposes which has led to a high potential for future research [17].

To conclude the methodology, current XAI research techniques revolve more around a development framework that usually consists of a combination of many post-hoc and intrinsic models that could show a substantial explanation for some user-specific tasks [15]. However, these techniques are not well-versed and have not been tested in different scenarios. Thus, it still raises questions about the selection of these XAI techniques for a given problem and there is no method by which one model can be preferred over the other. The selection process for the XAI process more or less depends on the objective through the explanations and results required and is mostly initiated by the experts in the XAI field.

C. EVALUATION

The third part of designing an XAI model which is the evaluation of XAI techniques is constantly improving and new techniques are discovered, which creates difficulty in validating the performance of these techniques. To validate these techniques, DARPA has created some human-centric evaluation measures such as trust assessment, mental model (understandability), user satisfaction, and task performance as shown in Figure 5. [16] Firstly, the *trust assessment* of any XAI system is a measure of

effectiveness and can be evaluated using user knowledge, coverage, confidence, and extensive usage. This explanation of evaluation helps to understand the user's trust in these XAI systems. Secondly, *understandability* is derived by user understanding of systems such as feature importance-based classification, user interpretation of system and algorithms, model output and failure prediction for model evaluation. Thirdly, *user satisfaction* is measured through user-interview, questionnaires, and expert case studies. Lastly, *task performance* measures the improvement of users' decision-making and can be implemented by model accuracy, throughput, tuning and selection this implementation can be both subjective and objective in nature.

Though human-centric measures are normally applied to XAI evaluation, the potential drawbacks of human bias towards simpler explanations can result in a biased transparent system for other users. Nonetheless, there is no benchmark for comparing these evaluation metrics for both human-centric approaches. Moreover, ML models can be severely exposed to constant threats of adversarial attacks [4]. Hence, recent developments have led to a computer-centred approach due to the high complexity of human-centric evaluation measures which can be very difficult to implement.

Now, the computer-centred method can be broadly characterized into two types: *Interpretability* is a metric of the ability of the system to explain human terms and *fidelity* is the system's ability to explain its behaviour accurately using the surrogate model [11]. Since these methods heavily rely on a human-centric approach it can be said that these methods are only used after the use of the human-centred technique. There can be other sub-techniques in computer-centred explanation such as robustness and localization capability covered in the article [12]. However, most of these evaluations like performance, understandability, fidelity, and interpretability are targeted towards users who are experts.

Evaluation is still in the developing phase, thus has been less implemented in the real world and has a lot of future research potential. This sums up the state of the art in XAI, and to elaborate and discuss the implementation of the models in real-world scenarios, let's discuss applications of XAI in different commercial and industrial fields.

IV. APPLICATIONS

As the digitisation of the world progresses very fast and enormous amounts of data are being produced every single day, Artificial Intelligence has been introduced into many fields in order to make use of that data and give solutions to chronical problems. The list of applications is huge. Finance, Healthcare, Education, Marketing, Sales, Robotics, Manufacturing are only just a few to mention. However, as most of the Machine Learning algorithms and especially Deep Learning algorithms are seen as "black boxes" the need for XAI is growing. In this context there are still sectors that are reluctant to incorporate AI in their functions mainly due to the fact that an error made by an AI can lead to loss of human lives. These sectors include Healthcare, Defence and Autonomous Driving. The initial aim of this review was to focus on these three particular sectors due to their critical importance, but unfortunately we were not able to find any previous research on the Defence sector that is publicly available, and decided to cover Manufacturing instead. Therefore, the sectors that this review focuses on are Healthcare, Autonomous Driving and Manufacturing and are covered in this part.

A. XAI IN HEALTHCARE

Healthcare is one of the fields where AI was firstly applied. Nowadays, AI has expanded in a wide variety of applications in this field such as Clinical Decision Support (CDS), Hospital Management, Predictive Medicine, Drug Discovery, Patient Data and Risk Assessment [29,30]. Given the continuous spread of AI in healthcare, and the criticality of this particular field as the human life factor is present, the need for explainable and understandable by humans AI models is greater than ever. A review in the literature shows that we are thankfully moving towards that direction, as there has been a lot of related work and research for the sake of helping doctors to save lives.

Wen Loh et al. (2022) in [29] provide a thorough review of the application of XAI for healthcare over the last decade, where they surveyed multiple articles that covered the most popular XAI techniques used in healthcare, and also they give suggestions regarding the future of XAI in this field. According to the authors, one of the most common use cases of XAI in healthcare is for providing visual explanations by highlighting the most important parts of an image that affect the model's prediction. GradCAM provides heatmaps and is the most popular technique for this application. Their review concludes by mentioning that the SHAP technique is overall the most used XAI technique in healthcare, as it is used for a variety of purposes such as disease prediction and hospital management by identifying the most significant features in the prediction of a model.

According to the authors, the majority of use cases of XAI in healthcare is for Clinical Decision Support (CDS). XAI techniques such as SHAP and GradCAM are used for predicting various diseases or for supporting hospital management practices. In particular, a great portion of the latest research has been focusing on the recent pandemic of COVID-19 and on its prognosis. Hou and Gao (2021) in [31] developed a platform that can "distinguish COVID-19

pneumonia from non-COVID-19 pneumonia". They feed image data such as chest X-rays into a deep convolutional neural network and use GradCAM to explain their models predictions.

A lot of work has also been carried out around heart disease detection. Dave et al. (2020) and Bahani et al. (2021) worked on the diagnosis of heart disease. Dave et al. in [32] conducted a study which includes many examples on predicting heart disease and supporting their results using various XAI feature-based techniques such as SHAP and LIME, and example-based techniques such as Counterfactuals and Anchors. The SHAP technique was quite successful in both local and global explanations. The conclusion from this study according to the authors is that all techniques that were used, help in showing which are the most influencing features for a model's prediction for heart disease. Bahani et al. in [30] propose a fuzzy clustering and linguistic modifiers-based method for creating a system that learns fuzzy classification for heart disease diagnosis. Their model was able to provide "transparent linguistic knowledge base" explanations, and also maintain a good trade-off between the interpretability and the accuracy. According to the authors, the end users can benefit from their proposed method as the knowledge base can detect problems within the data, and also optimise their models. Lastly, the knowledge base can help also in the decision-making stage.

Cancer is a disease that has drawn a lot of attention from the researchers and significant advancements have been made in cancer detection using XAI. Barata et al. (2020) in [33] developed a computer aided diagnostic system for detecting skin cancer. Their system is a combination of CNNs and GradCAM that gives visualisation explanations of the most influential region of an image and includes medical data and knowledge as it tries to mimic the hierarchical organisation of skin tumour as proposed by dermatologists. The results show that their model was able to achieve a competitive performance specially when incorporated the hierarchical decision process, and to finally give a skin cancer diagnosis by identifying it in the images. Another very common type of cancer is breast cancer and according to Binder et al. (2021) in [34] there are particular features of tumour pathology that need more study and attention. They propose a method for detecting breast cancer that combines multiple molecular and morphological features, and they make use of the LRP technique for providing explanations of their model's predictions. This particular method can identify cancer cells on historical images with high spatial resolution, and a variety of other molecular features and also rank them according to tumour pathology. According to the authors, their method cannot replace the Biochemistry science because the predictions made do not show the exact location of the molecular features in the images, but it can help to interpret clinal data and produce hypotheses on the significance of the features.

XAI has also been introduced to the concept of drug discovery. Jimenez-Luna, Grisoni and Schneider (2020) in their thorough review of drug discovery with XAI [35], discuss about the need for explainable methods and what exactly is expected from XAI to bring to drug discovery. According to the authors, it is needed to know how the model came to a prediction, the reason why this prediction is acceptable, assist humans in decision-making and finally estimate how reliable the prediction is. The design of new drugs is a quite complex process with many challenges around molecular pathology, and XAI can help to deal with some of these challenges. In particular, XAI can "take informed action while simultaneously considering medicinal chemistry knowledge".

Another important application of XAI in healthcare other that the CDS is for hospital management and for ensuring normal operation of the hospitals and clinics. Making use of the available resources efficiently, and identifying patients of high risk of reattending the emergency department early after their discharge, is of significant importance. In this context, Chmiel et al. (2021) in [36] and Lo et al. (2021) in [37] developed methods for identifying patients of high risk of early readmission. The method proposed by Chmiel et al. is able to identify the most relevant features of readmission risk which in this case is the medical history of the patient. Lo et al. focus on developing a method for identifying the risk of 14-day readmission of patients. According to the authors hospital readmissions can have multiple negative impacts both on the health of patients and to the hospital reputation as well as the medical costs for the hospital. Therefore, the identification of these patients can prevent these issues. Both of works used the SHAP XAI technique for explaining their models' decisions.

B. XAI IN AUTONOMOUS DRIVING

As in the Healthcare sector, XAI in Autonomous Driving is equally important as false decisions made by Autonomous Vehicles (AVs) can cost human lives. The deployment of AVs on roads can have multiple benefits [38] such as reduced traffic congestion, increased safety, lower carbon emissions, and they can be convenient to elderly people and those with kinetic problems. However, people are cautious towards the AVs due to the lack of explainability in their actions, and as more accidents occur involving AVs, their trust gets even more damaged. Therefore, there is a high need for explanations in order to increase the acceptance of AI in this field.

Atakishiyev et al. (2022) in [38], propose an XAI framework as seen in figure 1, that combines legal and social requirements for the development of XAI approaches in Autonomous Vehicles (AVs). Their proposed framework seems to be quite promising as it tries to explain the rationale behind the actions of AVs, while at the same time confirming safety and regulatory compliance. To achieve this, their proposed framework incorporates three main components. The first component is a Control System, that continuously maps all instances of a sensed environment to a set of corresponding actions that an AV can take. The second component is the Safety-Regulatory Compliance component, which is based on standards and certifies the safety of the Control System. This is achieved by taking into account results from software simulations, and depending on a predefined threshold of compliance performance, compliance is confirmed, and the system can be deployed. The third component is the XAI component that provides the explanations for each action taken by the control system and therefore by the AV. According to the authors the explanations can vary from visual and textual or even be a combination of these.

The same authors, Atakishiyev et al. (2022) in [39], produced another paper where they applied their framework in a case study to showcase its role. The case study is a traffic scenario of an AV involved accident. The proposed framework seems to play an important role in the post-accident investigations, as by recording its actions and the explanations behind these actions, the investigators' life becomes much easier in understanding the cause of the accident, or who's fault is it in case there is another vehicle involved. This can then be useful for taking the appropriate actions in order to improve the problematic system an AV, if the results show that the AV took a wrong action and caused the accident.

Another method for visual type explanations was proposed by Bojarski et al. (2016) in [40]. Their proposed method called VisualBackProp as the name suggests is based on the backpropagation procedure on Convolutional Neural Networks (CNNs), and its main use is for supporting the decisions generated by the CNN. This is done by showing what parts of an image and which particular pixels contribute the most to the prediction, and in the context of autonomous driving this could be cars, or road lane markings and edges. The fact that the feature maps of deeper layers in a CNN have more relevant information compared to shallow layers, and especially the last layer which contributes the most in determining the output is the basic foundation of this method. According to the authors, their method combines only the deep layers with the most relevant information and the shallow ones since they have higher resolution.

Another type of explanation is counterfactual explanations. According to Jacob et al. (2022) in [41], explanations must show the region of interest in an image, but at the same time they should also show what exactly caused the model to get to that outcome decision. Most post-hoc methods used in computer vision fail to do this, as they focus only on the former. Counterfactual explanations are seen in recent research as a way to provide content-based explanations. According to the authors, a counterfactual explanation is a version of an input image with very small but meaningful changes, that changes the decision made by the model in such a way that contradicts its initial decision. In the context of autonomous driving, it is of great importance that we know what exactly caused the AV to take a particular action. For example, a vehicle stopped at a traffic light because of the red light, or it stopped because of a crossing pedestrian. What would happen if we modified the colour of the traffic light, or removed the pedestrian, or if added more pedestrians and cars. How would these modifications change the output decision of the model? Jacob et al. propose a model called STEring counterfactual EXplanations (STEEX) for providing meaningfully interpretable counterfactual explanations with semantics. Apart from generating counterfactuals STEEX can also allow the users to select which region of the image they want the model to focus on for generating explanations.

Rjoub et al. (2022) in [42] make use of Federated Deep Reinforcement Learning (FL) to build a trusted system for autonomous driving. FL is a new ML approach where the model is being trained by different devices which cooperate for this purpose. The authors have chosen this approach as it can help overcome issues related to the limitations of data that are required for this work. Their main goal is to "improve the effectiveness and trustworthiness of the trajectory decisions for newcomer AVs". For achieving this they make use of existing AVs data by extracting information such as the features that contribute most to their working function. Then the application of reinforcement learning limits challenges such as the need for continuous manual design of models according to scenario, and motion heuristics, as the AV can automatically learn how to make the best decisions using expert data derived from the most-trusted existing AVs. The authors employ techniques such as SHAP for evaluating the importance of each AV features in order to select the most trusted AVs to incorporate in the training. Their proposed method outperformed in terms of accuracy other methods such as Deep Q Network achieving an excellent 95% accuracy.

C. XAI IN MANUFACTURING

With the rise of the fourth industrial revolution the sector of manufacturing is undergoing a transition to automated processes with the use of AI. The list of benefits of implementing AI in manufacturing is large. However, the same "black box" issue is relevant to all applications where AI is incorporated, and manufacturing is no exemption. In terms of XAI, according to [44] most XAI methods found in the manufacturing sector are built into frameworks, as the XAI methods by themselves are not enough in terms of explanations.

In an attempt to provide transparent decisions and predictions of AI in manufacturing, the XMANAI (eXplainable MANufacturing Artificial Intelligence) project which is funded by the European Union (EU) comes to transform Europe's manufacturing activities. According to official documents from the EU [43], XAI and particularly the XMANAI project will help in multiple domains in the manufacturing processes, such as increased performance and productivity, improved quality of products and services, optimised production and achieving better resource management, and accurately predicting product demand. As a result, all the formers will also reduce the production and maintenance costs. The XMANAI project is a framework that takes into account all steps involved in an AI system, e.g., from obtaining the input data, processing and feature extraction, to the application of ML/DL algorithms in order to make predictions, and XAI models for providing explanations, and allows all relevant stakeholders to collaboratively work by sharing data and AI models.

Meister et al. (2021) in [45] develop a method that can increase the inspection efficiency of components and therefore reduce the manufacturing time. In particular, their research focuses on the automated classification of defects of manufactured fibre layup which is used in aviation. According to the authors, as the manual inspection of such composite components is very time consuming with 50% of the total manufacturing time spent for this activity, the automated classification by AI can be of great help. The authors compare 20 different XAI techniques in collaboration with CNNs, on images with various defect types such as gaps, wrinkles, and twists and images with no defects. They concluded that the DeepSHAP and Smooth Integrated Gradients (IG) methods were the most appropriate for this process. Their model achieved the impressive 99.29% accuracy while at the same time the XAI techniques used helped in highlighting the most significant and influential parts of the images.

Yoo and Kang (2021) in [46] developed a method for cost estimation in manufacturing and the visualisation of machining features. Their method aims to estimate the manufacturing costs for online manufacturing platforms. In such platforms the end-user shares a Computer-Aided Design (CAD) model that wants to build and gets a quotation. According to the authors there are particular features of a CAD model that increase the manufacturing costs, and their identification is important for making accurate cost predictions. This proposed method provides information and guides the designers in meeting their cost targets and design criteria while still on the initial stages of product development. A CNN with Grad-CAM is utilised for visualising the CAD features that increase the manufacturing costs. Their model advises the designers how to modify the design in order to reduce the costs and meet their target and also can identify the difficulty of the machining process.

Brito et al. (2021) in [47] provide a new XAI approach for detecting and diagnosing faults in industrial rotating machinery. Rotating machinery is widely used in industrial applications and the detection of faults is important for taking the appropriate actions to mitigate the disruptions in manufacturing processes. The proposed method extracts features from vibration signals that give information about the dynamic behaviour of the equipment, detects faults based on anomaly detection algorithms, and uses the SHAP and Local-DIFFI XAI techniques to interpret the models predictions by highlighting the most relevant features that caused the fault. The authors with their proposed methodology were able to achieve impressively high accuracies of >99%, provide explainability, and use it for diagnosing the fault. In terms of XAI, the two techniques used gave very similar results, however Local-DIFFI showed a much better computational performance.

In terms of future work in XAI in manufacturing, Mugurusi and Oluka (2021) in [48] attempt to highlight the importance of XAI in the Supply Chain Management (SCM). According to the authors the spread of AI in this field is significant as the supply chains are shifting to data and information management from physical management like warehouses and transport as AI has the potential to improve several activities. However, their paper concludes that although there is evidence that XAI will help in the chain supply the level of development of XAI in this field is low and therefore more attention is needed.

V. EXPLAINABLE AI IN BUSINESS

A. BACKGROUND

For profit organizations, the need for an explanation from a machine is derived from a variety of factors such as product quality, monitoring operational and manufacturing needs, the compliance requirements for standards imposed by the governments and many other factors. As per Vermeire, T et al. [49], in most cases, the

Explainability is centred for the Machine Learning engineers who have technical aspects and not the business or end users who are more relevant when it comes to an end-product. Furthermore, the technical implementors use explainability for debugging the relationship between model parameters and the output yielded. The methodology for enhancing the transparency of ML models and therefore of the end-products, for stakeholders is equally if not more, essential.

Some of the generic approaches towards providing an explainability are described below [49]:

1) PUBLISHING ALGORITHM CODES

Some of the advantages of working with the software are to easily publish the source codes to a public platform. This will make sure the adopted algorithm will justify itself in those matters.

2) DISCLOSE THE PROPERTIES OF ALGORITHM

In addition, it would make more sense to disclose the parameter of the ML model(s), the training and testing methodologies with transparent results, along with the datasets and external tools used for the conclusion.

3) ADOPT A GENERAL METHODOLOGY

The gap between the stakeholders and the techno-savvy people cannot be bridged with just the above two aspects. Non-technical people will still struggle to grasp the technical details and hence there should be a common methodology that should work for both groups. According to U. Bhatt et al. [50], most of the past research was focused on finding new approaches to XAI that caters to developer's requirements without factoring stakeholder's requirements. From stakeholder's viewpoint, these requirements are in the forms of interests, goals, expectations, needs, and demands regarding ML systems are defined as "Stakeholders desiderata". Nowadays there is an increased amount of latest research that can be found that does not overlook the stakeholders' desiderata while exploring this space. Some of these are covered in the upcoming sections.

B. NEED FOR EXPLAINABILITY IN BUSINESS

This section provides an overview for the requirements of the XAI defined in [49] based on manual interviews taken with the technical implementation team who are key part of this overall structure.

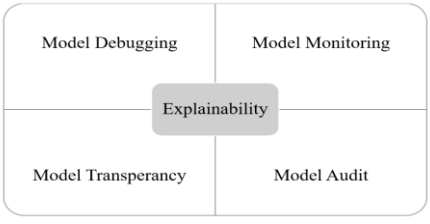

FIGURE 6. Explainability needs in business from a technical implementation viewpoint.

Model debugging can be used as an extraordinary tool for the ML model where the complexity is astronomical, and because it would take huge manual efforts and resources to optimize and explain it. If there was a meta-mode that can provide ML engineers with the capability of self-debugging in such scenarios, the deployment process can be expedited or prevented from possible execution bottlenecks.

Model Monitoring can be challenging especially when there are variations in terms of the drifts in one or more features after the deployment. Real-world scenarios often bring this deviation and constant monitoring which explains what and where the impact of these variations will help the technical team to apply robust solutions from time to time. This ends up enhancing the model's flexibility.

Model Transparency is undoubtedly a key aspect when it comes to regulatory compliance while and after the deployment of the ML mode. This also should aid the end users who are directly affected by the decision of the model's prediction. User manuals and public documentation will be easy if the model's workings can be explained in non-technical language for the general audience.

Model Audit is an essential step that comes prior to the deployment as part of an internal audit where the model is audited thoroughly for certain regulatory compliances such as security, sustainability, and stability. All three criteria are executed for small and big changes in the output of the model and then model is validated against the expectations of the outcome and an ideal behaviour.

C. DEFINING STAKEHOLDER'S DESIDERATA AND CRITERIA

First, from Markus Langer et al. [51] we define the categories of stakeholders to understand the problem space easily.

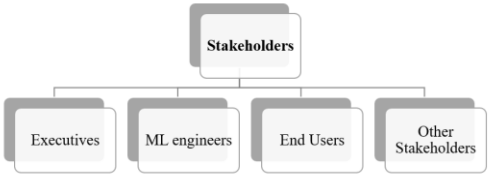

FIGURE 7. Stakeholders for which the desiderata are to be defined.

Executives: These people are decision-makers who drive their businesses for achieving success. Also, they are required to be making their business more compliant with some of the government rules and regulations.

ML Engineers: Individuals or groups of researchers and data scientists, who design, train, and implement ML models for the technological gains in the market. They use explainability to debug and prove why they work or behave in a certain way.

End Users: These are consumers who are affected by the results of the output of ML models. Explainability serves them with the sense of trust and reliability of using the product or technology or services and explains the behaviour in a much more non-technical way than the ML engineers would expect.

Other Stakeholders: This group can contain many different types of stakeholders such as auditors, safety officials, regulators, analysts, domain experts of the fields etcetera, who are required to understand the possible methods and techniques for explaining the ML model, the possible depth of the research area, and the relationship between the intuition and explanation from the model's output.

D. OVERVIEW OF STAKEHOLDER'S ROLES

For the above stakeholders, the defined desiderata can be summarised as follows in Markus Langer [2.10].

TABLE III
STAKEHOLDERS DESIDERATA

Desideratum	Business Role	Stakeholder
Acceptance	Deployer, Regulator	Executives, ML Engineers, Others
Accountability	Regulator	Executives, Others
Accuracy	Developer	ML Engineers
Autonomy	User	End Users
Confidence	User	End Users
Controllability	User	End Users
Debuggability	Developer	ML Engineers
Education	User	End Users
Effectiveness	Developer, User	ML Engineers, End Users
Efficiency	Developer, User	ML Engineers, End Users
Fairness	Affected, Regulator	Others
Informed Consent	Affected, Regulator	Others
Legal Compliance	Deployer	ML Engineers
Morality/Ethics	Affected, Regulator	Others
Performance	Developer	ML Engineers
Privacy	User	End Users, Others
Responsibility	Regulator	Executives, Others
Robustness	Developer	ML Engineers
Safety	Deployer, User	ML Engineers, End Users
Satisfaction	User	End Users
Science	User	End Users
Security	All	Executives, ML Engineers, End Users, Others
Transferability	Developer	ML Engineers
Transparency	Regulator	Executives, Others
Trust	User, Deployer	ML Engineers
Trustworthiness	Regulator	Executives, ML Engineers, End Users, Others
Usability	User	End Users
Usefulness	User	End Users
Verification	Developer	ML Engineers

The connectivity between stakeholders and their possible role in the business can be defined as per the following relationship suggested by Christian Meske et al.[52]

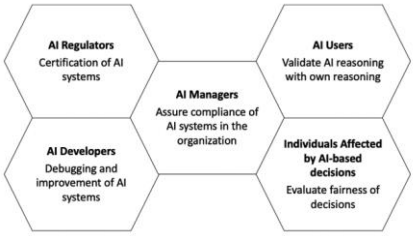

FIGURE 8. Stakeholders' various business roles and description of the responsibilities as defined in [52]

E. BUSINESS IMPLEMENTATION METHODOLOGIES FOR XAI

An overview of how XAI can enhance the overall business processes by adding explainability and interpretability in parallel to the decision-making can be illustrated in Figure 9 from C. Ouyang et al. [53]

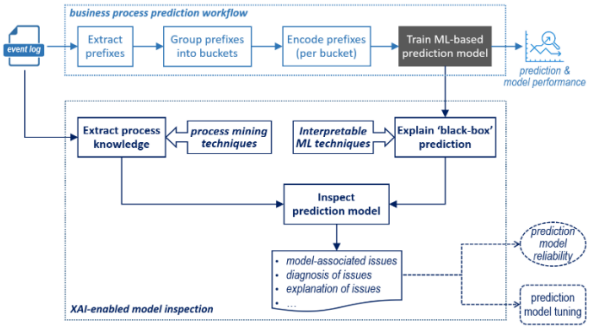

FIGURE 9. XAI embedment in business process sourced from [53]

With compared to the algorithmic and technological advancement in the field of XAI, the current literature still lacks some research perspectives to serve the explainability on the stockholders' side according to T Vermeire et al. [49]. There can be a simple naïve survey approach employed in a typical manner where the implementors interview different teams concerning the implementation and produce the customized solution.

One of the promising proposed approaches found [50] involves matching the explainability needs and explanation method properties. By conjoining these two aspects, we have a template containing formal documentation. This documentation consists of different Identity Cards or ID cards describing the explanation methods in detail.

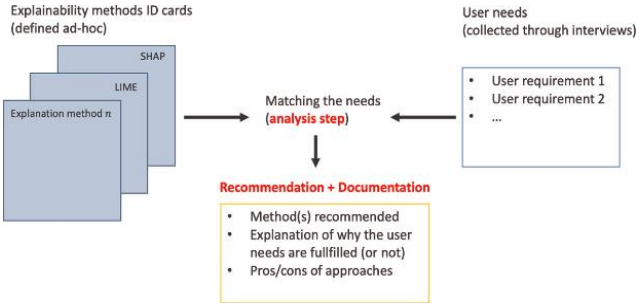

FIGURE 10. XAI implementation methodology proposed in T Vermeire et al. [49]

The ID card stack can be modelled as per shown in Figure 11. This is usually prepared with the help research community by the technical team consisting of ML engineers and data scientists based on available information or development depending upon how challenging the implementation is.

To document the user needs, there is a separate stack of cards based on different user requirements as shown in Figure 12. This is conducted with the help of individual stakeholders by reviewing their existing processes and interviewing them for investing the scope of their need for the explanation in their relevant areas.

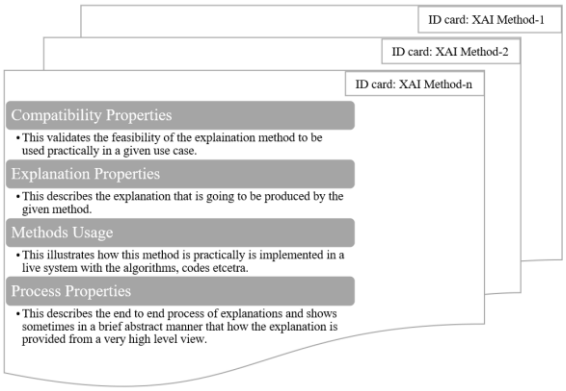

FIGURE 11. Explainability method ID cards with properties

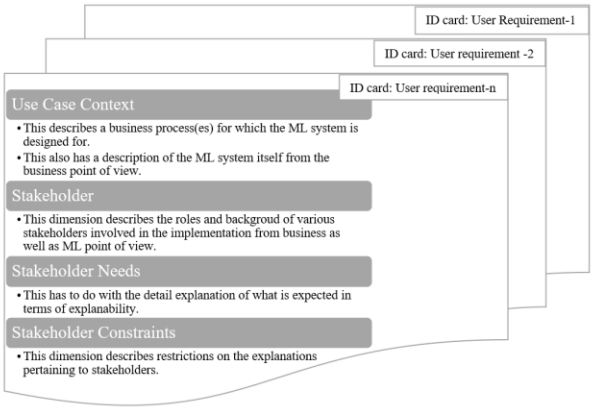

FIGURE 12. User requirements ID cards with dimensions

Once the documentation is completed meeting both the ends mentioned in the approach suggested in Figure 10, the information collection phase can be concluded, with a formal draft prepared to maintain and to refer from during the scope of the implementation.

F. CURRENT SCOPE OF XAI IN BUSINESS

The algorithmic and hardware advancement from the past decade has enabled businesses to get their business processes automated at every level of execution and operations. According to a software technology giant organization, Birla Soft [54], below are the high-level business use cases embedded with the XIA methodologies. As per the author of this article, businesses can get benefits from XIA by gaining the following positives:

1) Reducing the costs of mistakes
2) Reducing the impact of biasness of models
3) Increased responsibility and accountability
4) Code Confidence
5) Code compliance

FIGURE 13. XAI Business use cases as mentioned in [54]

VI. EXPERIMENT- IMPLEMENTING EXPLAINABILITY FOR MNIST IMAGES

MNIST is a hugely popular dataset in the deep learning world, produced by Yann Lecun [55] containing grayscale images of handwritten English digits from 0 to 9. The dataset in this demo is taken from the dataset catalogue of the TensorFlow library [56] which is a replica of the original MNIST dataset. Below are some sample images from the mentioned dataset.

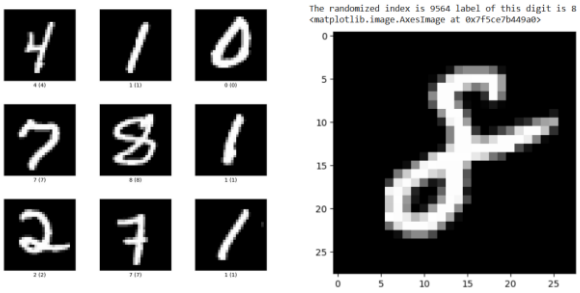

FIGURE 14. Sample from TensorFlow MNIST dataset to be considered for the explanation using LIME (left) and MNIST dataset sample images and their respective digits from TensorFlow dataset catalogue (right)

1) Train a Convolutional Neural Network for using the training set of 60000 images and measuring the accuracy of the trained model using the testing set of 10000 images
2) Implement the LIME algorithm to justify the prediction for a randomly selected sample from the testing set of 10000 images.
3) Analyzing and discussing the results

A. CNN TRAINING

Before we supply the training data to CNN, there is some pre-processing done considering conventional neural network rules such as image normalization and converting the grey-scale images to multidimensional images. The neural network architecture has the following layers as per mentioned in Fig 15. The training is done in five epochs. The derived testing accuracy is 0.9877 with a testing loss of 0.0582 and a testing accuracy of 0.9811 after the final epoch.

B. IMPLEMENTATION OF LIME

A pseudorandom sample of index 9564 was taken from the test image set which was truly labelled as digit ‘8’. The handwritten image is shown in Figure 14.

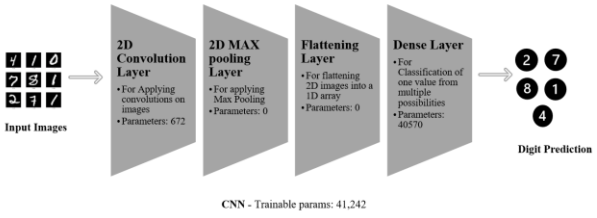

FIGURE 15. CNN architecture implemented using Keras for MNIST dataset images to digit prediction.

We have used below essential Python libraries [57] for LIME explanation implementation:
1) *lime*: for *LimeImageExplainer* object
2) *skimage* (Scikit image) [58]: to get the segment from images- shapes curves etcetera from *mark_boundaries* class
3) *matplotlib* [59]*:* to plot image results

LimeImageExplainer class takes our CNN model’s prediction method as input, along with various hyperparameters and the testing sample and gives out an *explanation* instance which can be used in analyzing results and providing the justification.

C. ANALYSIS AND DISCUSSION FOR EXPLANATION

Our analysis for explanation is done for two major aspects:
1) Decision boundaries are given by LIME algorithms for the given prediction done by the model. In Figure 13, the left image shows the decision boundary denoted in yellow drawn by the LIME algorithm.

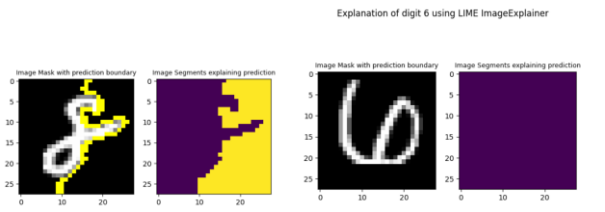

FIGURE 16. Explanation by marking Decision boundary and by Image Segmentation(left) and ambiguity of explanations(right)

Using this, the decision was taken by our CNN that the digit is ‘8’ out of ten possible numbers. This is done by deriving a superpixel for a given set of pixels.
2) Image Segmentation: Using the segmentations, LIME helps us derive areas of interest by which a justification can be provided. In our example for digit ‘8’, in Figure 13, the right image, there are two different areas which are separated by the decision boundary marker in the above aspect. The CNN, reportedly proven as an expert of image classifier learns from these areas and its separations to predict.

The data set also contains several ambiguous handwriting images, which is not conclusive enough for CNN to have a clear decision. Such examples are shown in Figure 14. As the justifications provided by LIME are blank in both above-discussed aspects, this ambiguity can also be detected clearly by the algorithm.

Few more examples of more detailed detections can be seen in below images combined under Fig 16.

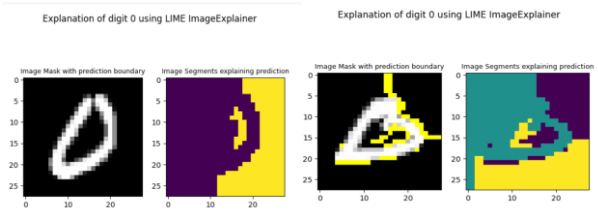

FIGURE 17. Different boundary markers and Image segmentations for handwritten digit '0'

The implementation code for the above demo is available at [60]. From a production viewpoint, this can be extended to more complex real-life scenarios of interpretation and justification in different areas of applications mentioned in the former section of this document. Those fields deal with different natures of input data such as images from chest X-rays, CT Scans, Microscopic images, real-time camera feed, aerial surveillance and so on. Although, so far, we covered a tiny experiment on how the explainability works at its root using an extremely simple dataset, empirically, detailed images with more colour channels, with more pixels and more capable and sophisticated several layers deep Neural Network architecture, shall generate more accurate results for those images. In addition to this, cutting-edge hardware configuration can help achieve high-level accuracy for production-quality implementations.

VII. FUTURE OF XAI

With the advancement of Deep learning in the past decade, more complex problems can be solved by ML models with better accuracy. This is attributed to the increment of the complexity of models, and they become more sensitive and susceptible to the variety and diversity of the problem space. Their size in terms of trainable parameters also increases with each generation. This poses a unique challenge of explainability in specifically decision-sensitive fields such as medicine, law, finance, defence etc. The decisions in many areas under these fields are related to many socioeconomic issues and touch on various aspects of moral compliance.

In a meta-survey done by [61], suggest possible research direction of the future for XAI given the current challenges faced in existing methodologies. Most of the current literature mentioned discusses the improvement in below aspects:

1) Addressing the challenges in existing XAI methods
2) Contrastive and counterfactual explanations
3) Communicating uncertainties
4) Time constraints
5) Natural Language Generation
6) Reproducibility

In a broader and more general sense, according to PwC survey on XAI from 2017 [62], roughly 67% of corporate leaders around the world believed that AI and automation will impact stakeholders' trust, largely in a negative manner in the next few years. This sounds alarming that we are almost a few years from this survey now and there is a growing need for explainability than ever in the AI realm. Furthermore, the same report mentions that the market size of AI is already $15 trillion which clearly suggests more and more black box decisions are being taken without non-significant justification considering the need of stakeholders. This will open another window of burning need opportunity for Explainable AI.

VIII. CONCLUSION

Explainable Artificial Intelligence – XAI is a field of study that focuses on developing strategies based on AI to make AI more transparent, understandable, and trustworthy to humans. Current algorithms are used for a variety of decision-critical applications because of significant advancements in AI over the past few years. XAI has been proposed to meet this demand for greater AI transparency and hasten the adoption of AI in significant domains. However, model complexity has been increased and opaque black-box AI models have been developed and employed by many researchers. There are numerous studies of XAI subjects in the literature that have identified issues and potential directions for XAI research in the future. As we provide a general overview of current challenges and potential research, the implementation techniques are still scattered and under improvement given that the field of AI progresses by leaps and bounds. In addition, this paper provides an elaborated insight into the current and future status of XAI and suggests that it is a promising field of research with the potential to make AI systems more trustworthy and beneficial to society. With the ever-growing demand for automation and AI systems, we infer that the need for XAI will only grow over time.

REFERENCES

[1] A. Holzinger, R. Goebel, R. Fong, T. Moon, · Klaus-Robert Müller, and W. Samek, "xxAI-Beyond Explainable AI," 2020. [Online]. Available: https://link.springer.com/bookseries/1244

[2] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges."

[3] H. K. Dam, T. Tran, and A. Ghose, "Explainable software analytics," in *Proceedings - International Conference on Software Engineering*, IEEE Computer Society, May 2018, pp. 53–56. doi: 10.1145/3183399.3183424.

[4] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[5] S. Richardson, "Exposing the many biases in machine learning," *Business Information Review*, vol. 39, no. 3, pp. 82–89, Sep. 2022, doi: 10.1177/02663821221121024.

[6] L. Hofeditz, S. Clausen, A. Rieß, M. Mirbabaie, and S. Stieglitz, "Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring," *Electronic Markets*, vol. 32, no. 4, pp. 2207–2233, Dec. 2022, doi: 10.1007/s12525-022-00600-9.

[7] K. Werder, B. Ramesh, and R. S. Zhang, "Establishing Data Provenance for Responsible Artificial Intelligence Systems," *ACM Trans Manag Inf Syst*, vol. 13, no. 2, Jun. 2022, doi: 10.1145/3503488.

[8] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, "Building Ethics into Artificial Intelligence," Dec. 2018, [Online]. Available: http://arxiv.org/abs/1812.02953

[9] D. Saraswat et al., "Explainable AI for Healthcare 5.0: Opportunities and Challenges," IEEE Access, pp. 1–1, 2022, doi: https://doi.org/10.1109/access.2022.3197671.

[10] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, "Survey of Explainable AI Techniques in Healthcare," Sensors, vol. 23, no. 2, p. 634, Jan. 2023, doi: https://doi.org/10.3390/s23020634.

[11] P. Lopes, E. Silva, C. Braga, T. Oliveira, and L. Rosado, "XAI Systems Evaluation: A Review of Human and Computer-Centred Methods," Applied Sciences, vol. 12, no. 19, p. 9423, Sep. 2022, doi: https://doi.org/10.3390/app12199423.

[12] A. Rawal, J. Mccoy, D. B. Rawat, B. Sadler, and R. Amant, "Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges and Perspectives," IEEE Transactions on Artificial Intelligence, pp. 1–1, 2021, doi: https://doi.org/10.1109/tai.2021.3133846.

[13] A. M. Antoniadi et al., "Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review," Applied Sciences, vol. 11, no. 11, p. 5088, May 2021, doi: https://doi.org/10.3390/app11115088.

[14] E. M. Kenny and M. T. Keane, "Explaining Deep Learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI," Knowledge-Based Systems, vol. 233, p. 107530, Dec. 2021, doi: https://doi.org/10.1016/j.knosys.2021.107530.

[15] M. T. Keane and E. M. Kenny, "How Case-Based Reasoning Explains Neural Networks: A Theoretical Analysis of XAI Using Post-Hoc Explanation-by-Example from a Survey of ANN-CBR Twin-Systems," Case-Based Reasoning Research and Development, pp. 155–171, 2019, doi: https://doi.org/10.1007/978-3-030-29249-2_11.

[16] X. Li, et al.,"A Survey of Data-Driven and Knowledge-Aware eXplainable AI" in IEEE Transactions on Knowledge & Data Engineering, vol. 34, no. 01, pp. 29-49, 2022. doi: 10.1109/TKDE.2020.2983930.

[17] T. Clement, N. Kemmerzell, M. Abdelaal, and M. Amberg, "XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process," Machine Learning and Knowledge Extraction, vol. 5, no. 1, pp. 78–108, Mar. 2023, doi: https://doi.org/10.3390/make5010006.

[18] E. M. Kenny, E. D. Delaney, D. Greene, and M. T. Keane, "Post-hoc Explanation Options for XAI in Deep Learning: The Insight Centre for Data Analytics Perspective," Pattern Recognition. ICPR International Workshops and Challenges, pp. 20–34, 2021, doi: https://doi.org/10.1007/978-3-030-68796-0_2.

[19] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," arXiv.org, 2015. https://arxiv.org/abs/1512.04150

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," International Journal of Computer Vision, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: https://doi.org/10.1007/s11263-019-01228-7.

[21] J. Darias, B. Díaz-Agudo, and J. Recio-Garcia, "A Systematic Review on Model-agnostic XAI Libraries." Accessed: Apr. 03, 2023. [Online]. Available: https://ceur-ws.org/Vol-3017/96.pdf

[22] F. Camastra, A. Ciaramella, S. Sposato, and A. Staiano, "A Fuzzy Rule Base Minimization Perspective in XAI." Accessed: Apr. 07, 2023. [Online]. Available: https://ceur-ws.org/Vol-3074/paper12.pdf

[23] G. Vilone and L. Longo, "Classification of Explainable Artificial Intelligence Methods through Their Output Formats," Machine Learning and Knowledge Extraction, vol. 3, no. 3, pp. 615–661, Aug. 2021, doi: https://doi.org/10.3390/make3030032.

[24] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," *arXiv:1704.02685 [cs]*, Oct. 2019, Available: https://arxiv.org/abs/1704.02685

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *arXiv.org*, 2016. https://arxiv.org/abs/1602.04938

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, Available: https://ojs.aaai.org/index.php/AAAI/article/view/11491

[27] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *arXiv:1705.07874 [cs, stat]*, Nov. 2017, Available: https://arxiv.org/abs/1705.07874

[28] H. Chen, S. Lundberg, and S.-I. Lee, "Explaining Models by Propagating Shapley Values of Local Components," arXiv:1911.11888 [cs, stat], Nov. 2019, Accessed: Apr. 18, 2023. [Online]. Available: https://arxiv.org/abs/1911.11888

[29] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," Computer Methods and Programs in Biomedicine, vol. 226, p. 107161, Nov. 2022, doi: https://doi.org/10.1016/j.cmpb.2022.107161

[30] K. Bahani, M. Moujabbir, and M. Ramdani, "An accurate Fuzzy Rule-Based Classification Systems for heart disease diagnosis", Scientific African, p. e01019, Oct. 2021, doi: https://doi.org/10.1016/j.sciaf.2021.e01019

[31] J. Hou and T. Gao, "Explainable DCNN based chest X-ray image analysis and classification for COVID-19 pneumonia detection", Scientific Reports, vol. 11, no. 1, Aug. 2021, doi: https://doi.org/10.1038/s41598-021-95680-6

[32] D. Dave, H. Naik, S. Singhal, and P. Patel, "Explainable AI meets Healthcare: A Study on Heart Disease Dataset", arXiv:2011.03195 [cs], Nov. 2020, Available: https://arxiv.org/abs/2011.03195#

[33] C. Barata, M. E. Celebi, and J. S. Marques, "Explainable skin lesion diagnosis using taxonomies", Pattern Recognition, vol. 110, p. 107413, Feb. 2021, doi: https://doi.org/10.1016/j.patcog.2020.107413

[34] A. Binder, M. Bockmayr, M. Hagele, S. Wienert, D. Heim, K. Hellweg, M. Ishii, A. Stenzinger, A. Hocke, C. Denkert, K. Muller and F. Klauschen, "Morphological and molecular breast cancer profiling through explainable machine learning", Nature Machine Intelligence, vol. 3, no. 4, pp. 355–366, Mar. 2021, doi: https://doi.org/10.1038/s42256-021-00303-4

[35] J. Jiménez-Luna, F. Grisoni, and G. Schneider, "Drug discovery with explainable artificial intelligence," *Nature Machine Intelligence*, vol. 2, no. 10, pp. 573–584, Oct. 2020, doi: https://doi.org/10.1038/s42256-020-00236-4

[36] F. P. Chmiel, D. K. Burns, M. Azor, F. Borca, M. J. Boniface, Z. D. Zlatev, N. M. White, T. W. V. Daniels and M. Kiuber, "Using explainable machine learning to identify patients at risk of reattendance at discharge from emergency departments", Scientific Reports, vol. 11, no. 1, Nov. 2021, doi: https://doi.org/10.1038/s41598-021-00937-9

[37] Y.-T. Lo, J. C. Liao, M.-H. Chen, C.-M. Chang, and C.-T. Li, "Predictive modeling for 14-day unplanned hospital readmission risk by using machine learning algorithms", BMC Medical Informatics and Decision Making, vol. 21, no. 1, Oct. 2021, doi: https://doi.org/10.1186/s12911-021-01639-y

[38] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: An overview and guide for future research directions", arXiv:2112.11561 [cs], Apr. 2022, Available: https://arxiv.org/abs/2112.11561

[39] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Towards Safe, Explainable, and Regulated Autonomous Driving", arXiv:2111.10518 [cs], Apr. 2022, Available: https://arxiv.org/abs/2111.10518

[40] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, U. Muller, and K. Zieba, "VisualBackProp: visualizing CNNs for autonomous driving", 2016
[41] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, and M. Cord, "STEEX: Steering Counterfactual Explanations with Semantics", arXiv:2111.09094 [cs], Jul. 2022, Available: https://arxiv.org/abs/2111.09094
[42] G. Rjoub, J. Bentahar, and O. A. Wahab, "Explainable AI-based Federated Deep Reinforcement Learning for Trusted Autonomous Driving", 2022 International Wireless Communications and Mobile Computing (IWCMC), May 2022, doi: https://doi.org/10.1109/iwcmc55113.2022.9824617
[43] XMANAI Consortium, "XMANAI – Making AI Understandable", 2020, Available: https://ai4manufacturing.eu/
[44] G. Sofianidis, J. Rožanec, D. Mladenić, D. Kyriazis, and J. Stefan, "A Review of Explainable Artificial Intelligence in Manufacturing.", Jul. 2021, Available: https://arxiv.org/pdf/2107.02295.pdf
[45] S. Meister, M. Wermes, J. Stüve, and R. M. Groves, "Investigations on Explainable Artificial Intelligence methods for the deep learning classification of fibre layup defect in the automated composite manufacturing", Composites Part B: Engineering, vol. 224, p. 109160, Nov. 2021, doi: https://doi.org/10.1016/j.compositesb.2021.109160
[46] S. Yoo and N. Kang, "Explainable artificial intelligence for manufacturing cost estimation and machining feature visualization," Expert Systems with Applications, vol. 183, p. 115430, Nov. 2021, doi: https://doi.org/10.1016/j.eswa.2021.115430
[47] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. V. Duarte, "An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery," Mechanical Systems and Signal Processing, vol. 163, p. 108105, Jan. 2022, doi: https://doi.org/10.1016/j.ymssp.2021.108105
[48] G. Mugurusi and P. N. Oluka, "Towards Explainable Artificial Intelligence (XAI) in Supply Chain Management: A Typology and Research Agenda," Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems, pp. 32–38, 2021, doi: https://doi.org/10.1007/978-3-030-85910-7_4
[49] Vermeire, T., Laugel, T., Renard, X., Martens, D., Detyniecki, M. (2021). How to Choose an Explainability Method? Towards a Methodical Implementation of XAI in Practice. In:, et al. Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2021. Communications in Computer and Information Science, vol 1524. Springer, Cham. https://doi.org/10.1007/978-3-030-93736-2_39
[50] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley, "Explainable machine learning in deployment," arXiv.org, 10-Jul-2020. [Online]. Available: https://arxiv.org/abs/1909.06342
[51] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, Kevin 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,Volume 296,2021,103473,ISSN 0004-3702, https://doi.org/10.1016/j.artint.2021.103473 Available: https://www.sciencedirect.com/science/article/pii/S0004370221000242
[52] C. Ouyang, R. Sindhgatta, and C. Moreira, "Explainable AI enabled inspection of Business Process Prediction Models," arXiv.org, 16-Jul-2021. [Online]. Available: https://arxiv.org/abs/2107.09767
[53] "Demystifying explainable artificial intelligence: Benefits, use cases, and Models," Birlasoft. [Online]. Available: https://www.birlasoft.com/articles/demystifying-explainable-artificial-intelligence.
[54] "The mnist database," MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges. [Online]. Available: http://yann.lecun.com/exdb/mnist/
[55] "MNIST Tensorflow datasets," TensorFlow. [Online]. Available: https://www.tensorflow.org/datasets/catalog/mnist.
[56] Marcotcr, "MARCOTCR/Lime: Lime: Explaining the predictions of any machine learning classifier," GitHub. [Online]. Available: https://github.com/marcotcr/lime
[57] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu and the scikit-image contributors. scikit-image: Image processing in Python. PeerJ 2:e453 (2014) https://doi.org/10.7717/peerj.453
[58] "Visualization with python," Matplotlib. [Online]. Available: https://matplotlib.org/
[59] N. Patel, "Nikilkumarpatel_ee986_xai_lime_mnist_demo," Kaggle, 15-Apr-2023. [Online]. Available: https://www.kaggle.com/nikilpatel94/nikilkumarpatel-ee986-xai-lime-mnist-demo
[60] Waddah Saeed, Christian Omlin,Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, Knowledge-Based Systems,Volume 263,2023,110273,ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2023.110273.(https://www.sciencedirect.com/science/article/pii/S0950705123000230)
[61] PricewaterhouseCoopers, "Explainable AI," PwC. [Online]. Available: https://www.pwc.co.uk/xai